Homology modelling and protein engineering strategy of subtilases, the family of subtilisin-like serine proteinases

Siezen, Roland J.; Vos, Willem M. de; Leunissen, Jack A.M.; Dijkstra, Bauke W.

Published in:
%22Protein Engineering%2C Design and Selection%22

DOI:
10.1093/protein/4.7.719

Publication date:
1991

Protein Engineering vol.4 no.7 pp.719–737, 1991

RESEARCH REVIEW

Homology modelling and protein engineering strategy of subtilases, the family of subtilisin-like serine proteinases

Roland J.Siezen[1], Willem M.de Vos, Jack A.M.Leunissen[2] and Bauke W.Dijkstra[3]

Department of Biophysical Chemistry, NIZO, PO Box 20, 6710BA Ede, [2]CAOS/CAMM Center, University of Nijmegen, Toernooiveld, 6525ED Nijmegen and [3]Department of Chemical Physics, University of Groningen, Nijenborgh 16, 9747AG Groningen, The Netherlands

[1]To whom correspondence should be addressed

Subtilases are members of the family of subtilisin-like serine proteases. Presently, >50 subtilases are known, >40 of which with their complete amino acid sequences. We have compared these sequences and the available three-dimensional structures (subtilisin BPN′, subtilisin Carlsberg, thermitase and proteinase K). The mature enzymes contain up to 1775 residues, with N-terminal catalytic domains ranging from 268 to 511 residues, and signal and/or activation-peptides ranging from 27 to 280 residues. Several members contain C-terminal extensions, relative to the subtilisins, which display additional properties such as sequence repeats, processing sites and membrane anchor segments. Multiple sequence alignment of the N-terminal catalytic domains allows the definition of two main classes of subtilases. A structurally conserved framework of 191 core residues has been defined from a comparison of the four known three-dimensional structures. Eighteen of these core residues are highly conserved, nine of which are glycines. While the α-helix and β-sheet secondary structure elements show considerable sequence homology, this is less so for peptide loops that connect the core secondary structure elements. These loops can vary in length by >150 residues. While the core three-dimensional structure is conserved, insertions and deletions are preferentially confined to surface loops. From the known three-dimensional structures various predictions are made for the other subtilases concerning essential conserved residues, allowable amino acid substitutions, disulphide bonds, Ca^{2+}-binding sites, substrate-binding site residues, ionic and aromatic interactions, proteolytically susceptible surface loops, etc. These predictions form a basis for protein engineering of members of the subtilase family, for which no three-dimensional structure is known.

Key words: homology modelling/sequence alignment/serine proteinase/subtilase/subtilisin family

Introduction

The most extensive 'engineering' of proteins has occurred during the process of natural selection in the evolution of living organisms. Therefore, it seems appropriate to learn from nature and establish rules by the careful comparison of divergently evolved families of protein structures. This can be achieved through the multiple alignment of amino acid sequences to identify conserved residues, followed by superposition of known three-dimensional structures of family members to identify conserved conformations or motifs (Sutcliffe *et al.*, 1987a). Indeed, the comparison of three-dimensional structures and subsequent homology (or knowledge-based) modelling of different families of proteolytic enzymes, e.g. neutral proteases (Eijsink *et al.*, 1990; Signor *et al.*, 1990), aspartic proteases (Abad-Zapatero *et al.*, 1990; Blundell *et al.*, 1990; Weber, 1990), cysteine proteases (Bazan and Fletterick, 1988; Gorbalenya *et al.*, 1989) and trypsin-like serine proteases (Greer, 1981, 1990), has demonstrated that functionally significant regions of related proteins can be modelled with high accuracy.

Serine proteinases (EC 3.4.21.–) are of extremely widespread occurrence and diverse function. Although many distinct families of serine proteases seem to exist, the two best-studied ones are the (chymo)trypsin and subtilisin (EC 3.4.21.14) families. As shown below, these two families are often incorrectly refered to as the mammalian and microbial serine proteinases, or the eukaryotic and prokaryotic serine proteinases. These families are distinguished by a highly similar arrangement of catalytic His, Asp and Ser residues in radically different β/β (trypsin) and α/β (subtilisin) protein scaffolds. The well-known (chymo)trypsin family has numerous members (e.g. chymotrypsin, trypsin, elastase, plasmin, thrombin, kallikrein, factor IX, tPA), which are found in procaryotic and eucaryotic micro-organisms, invertebrates and vertebrates (Rogers, 1985; Irwin *et al.*, 1988; Greer, 1990). They exhibit widely varying degrees of sequence similarity, but display a high level of tertiary structure conservation of the catalytic domain (Greer, 1990).

Until recently, far fewer members of the subtilisin-like family were known and previously they appeared to be limited to micro-organisms. Only the three-dimensional structures of subtilisin BPN′/Novo from *Bacillus amyloliquefaciens* (Hirono *et al.*, 1984; McPhalen *et al.*, 1985; McPhalen and James, 1988; Bott *et al.*, 1988), subtilisin Carlsberg from *B.licheniformis* (Bode *et al.*, 1986, 1987; McPhalen and James, 1988), thermitase from *Thermoactinomyces vulgaris* (Gros *et al.*, 1989a,b, 1991; Teplyakov *et al.*, 1990) and proteinase K from *Tritirachium album* (Betzel *et al.*, 1988b,c), each consisting of ~275–280 residues, have been reported. The coordinates of an alkaline protease with extremely high pH optimum from *Bacillus alcalophilus* will soon be available (Sobek *et al.*, 1990). Their secondary structure elements and overall folding are very similar despite considerable differences in amino acid sequence. Subtilisins have considerable industrial importance as a protein-degrading component of washing powders; they have been extensively studied and engineered in the past decade to provide insight into both their stability and the mechanism and specificity of enzyme catalysis (reviewed by Wells and Estell, 1988).

In the past few years many new subtilisin-like serine proteinases have been found in archaea, bacteria, fungi, yeasts and even in higher eukaryotes (Table I). In this article we compare the amino acid sequences of all known subtilisin-like enzymes, which we propose to call 'subtilases'. From this multiple sequence alignment and the four known three-dimensional structures we have attempted to identify the essential framework or core structure of the catalytic domain of all subtilases, together with the variations that are allowed in the main-chain length and in the character of side chains. This information allows us to develop

Table I. The subtilase family of serine proteases

Organism	cDNA,gene	enzyme	acronym	cellular location	amino acids total/prepro/mature			signal peptide	membr. bound	C-term. proces.	C-term. repeats	3D
PROKARYOTES												
Bacteria: Gram-positive												
Bacillus subtilis 168	aprA	subtilisin I168, aprA	BSS168	extra	381	106	275	+	-	-	-	-
Bacillus amyloliquefaciens	apr	subtilisin BPN' (NOVO)	BASBPN	extra	382	107	275	+	-	-	-	+
Bacillus subtilis DY	-	subtilisin DY	BSSDY	extra	?	?	274	?	-	-	-	-
Bacillus licheniformis	+	subtilisin Carlsberg	BLSCAR	extra	379	105	274	+	-	-	-	+
Bacillus lentis	+	subtilisin 147	BLS147	extra	361	93	268	+	-	-	-	-
Bacillus alcalophilus PB92	+	subtilisin PB92	BAPB92	extra	380	111	269	+	-	-	-	-
Bacillus sp. DSM 4828	-	alkaline protease	BDSM48	extra	?	?	?	?	?	?	?	-
Bacillus YaB	ale	alkaline elastase YaB	BYSYAB	extra	378	110	268	+	-	-	-	-
Bacillus subtilis 168	epr	minor extracellular protease	BSEPR	extra	645	103	542	+	-	+	+	-
Bacillus subtilis	bpf	bacillopeptidase F	BSBPF	extra	1433	194	(1239)	+	-	+	-	-
Bacillus subtilis IFO3013	isp1	intracell. serine protease 1	BSISP1	intra	319	(0)	(319)	-	-	-	-	-
Bacillus subtilus A50	-	intracell. serine protease	BSIA50	intra	?	?	?	?	?	?	?	-
Bacillus thuringiensis	-	extracell. serine protease	BTFINI	extra	?	?	?	?	?	?	?	-
Bacillus cereus	-	extracell. serine protease	BCESPR	extra	?	?	?	?	?	?	?	-
Nocardiopsis dassonvillei (prasina)	-	alkaline serine protease	NDAPII	extra	?	?	?	?	?	?	?	-
Thermoactinomyces vulgaris	-	thermitase	TVTHER	extra	?	?	279	?	?	?	?	+
Enterococcus faecalis	cylA	cytolysin component A	EFCYLA	extra	412	?	?	+	-	-	-	-
Staphylococcus epidermidis	epiP	epidermin leader protease	SEEPIP	extra	461	?	?	+	?	?	-	-
Streptococcus pyogenes	scpA	C5a peptidase	SPSCPA	extra	1167	(31)	(1136)	+	+	-	+	-
Lactococcus lactis (cremoris) SK11	prtP	SK11 cell wall proteinase	LLSK11	extra	1962	187	(1775)	+	+	(+)	+	-
Bacteria: Gram-negative												
Dichelobacter nodosus	+	basic protease	DNEBPR	extra	603	132	344	+	-	+	-	-
Xanthomonas campestris	+	extracellular protease	XCEXPR	extra	580	(136)	(444)	+	-	?	-	-
Serratia marcescens IFO3046	+	extracellular serine protease	SMEXSP	extra	1045	27	(381)	+	-	(+)	-	-
Thermus aquaticus YT-1	pstI	aqualysin I	TAAQUA	extra	513	127	281	+	-	+	-	-
Thermus rT41A	+	T41A protease	TRT41A	extra	?	?	279	?	?	?	?	-
Vibrio alginolyticus	proA	protease A	VAPROA	extra	534	(141)	(393)	+	-	-	-	-
Streptomyces rutgersensis	-	proteinase D	SRESPD	extra	?	?	?	?	?	?	?	-
Archaea												
halophilic strain 172 P1	-	halophilic extracel.protease	ARB172	extra	?	?	?	?	?	?	?	-
Cyanobacteria												
Anabaena variabilis	prcA	Ca-dependent protease	AVPRCA	intra	>606	(191)	>415	-	?	?	?	-
LOWER EUKARYOTES												
Fungi												
Tritirachium album Limber	+	proteinase K	TAPROK	extra	384	105	279	+	-	-	-	+
Tritirachium album	+	proteinase R	TAPROR	extra	387	108	279	+	-	-	-	-
Tritirachium album	proT	proteinase T	TAPROT	extra	>293	?	281	?	-	-	-	-
Aspergillus oryzae ATCC20386	+	alkaline protease	AOALPR	extra	403	121	282	+	-	-	-	-
Malbranchea pulchella (sulfurea)	-	thermomycolin	MPTHMY	extra	?	?	?	?	?	?	?	-
Acremonium chrysogenum	alp	alkaline protease	ACALPR	extra	402	120	282	+	-	-	-	-
Yeasts												
Kluyveromyces lactis	kex1	Kex1 serine proteinase	KLKEX1	golgi	700	(126)	(574)	?	+	?	-	-
Saccharomyces cerevisiae	kex2	Kex2 serine proteinase	SCKEX2	golgi	814	(137)	(677)	+	+	?	+	-
Saccharamyces cerevisiae	prb1	protease B	SCPRB1	vacuole	635	280	(355)	+	-	(+)	-	-
Yarrowia lipolytica	xpr2	alk. extracellullar protease	YLXPR2	extra	454	157	297	+	-	-	-	-
HIGHER EUKARYOTES												
Worms												
Caenorhabditis elegans	bli4	cuticle protease	CEBLI4		>701	?	?	+	+	?	?	-
Insects												
Drosophila (fruit fly)	fur1	furin 1	DMFUR1		899	?	?	+	+	?	-	-
Drosophila (fruit fly)	fur2	furin 2	DMFUR2		>836	?	?	?	?	?	+	-
Plants												
Cucumis melo (melon)	-	cucumisin	CMCUCU		?	?	?	?	?	?	?	-
Mammals												
Human (also rat, mouse)	fur	furin	HSFURI		794	(107)	(687)	+	+	?	-	-
Human (also mouse)	+	insulinoma PC2 protease	HSIPC2	granule	638	(109)	(529)	+	-	?	-	-
Mouse	+	pituitary PC3 protease	MMPPC3		753	(110)	(643)	+	-	?	-	-
Human	+	tripeptidyl peptidase II	HSTPP	intra	>1196	?	?	?	-	-	-	-

Table I. continued

References to amino acid sequences (GenBank™/EMBL Data Bank accession numbers are shown in brackets):

ARB172 Kamekura,M. and Seno,Y. (1990) *Biochem. Cell Biol.*, **68**, 352–359 (amino acid sequencing of mature protease residues 1–35; residue 19 not determined).

BSS168 Stahl,M.L. and Ferrari,E. (1984) *J. Bacteriol.*, **158**, 411–418 (K01988). Yoshimoto,T., Oyama,H., Honda,T., Tone,H., Takeshita,T., Kamiyama,T. and Tsuru,D. (1988) *J. Biochem.*, **103**, 1060–1065 (the mature subtilisin from *B.subtilis* var. *amylosacchariticus* differs in having T130S and T162S). Svendsen,I., Genov,N. and Idakieva,K. (1986) *FEBS Lett.*, **196**, 228–232 (PIR A23624; amino acid sequencing; the mature alkaline mesentericopeptidase from *B.mesentericus* differs in having S85A, A88S, S89A, S183A and N259S).

BASBPN Wells,J.A., Ferrari,E., Henner,D.J., Estell,D.A. and Chen,E.Y. (1983) *Nucl. Acids Res.*, **11**, 7911–7925 (X00165). Vasantha,N., Thompson,L.D., Rhodes,C., Banner,C., Nagle,J. and Filpula,D. (1984) *J. Bacteriol*, **159**, 811–819 (K02496).

BSSDY Nedkov,P., Oberthur,W. and Braunitzer,G. (1983) *Hoppe-Seyler's Z. Physiol. Chem.*, **364**, 1537–1540 (PIR A00969; amino acid sequencing).

BLSCAR Jacobs,M., Eliasson,M., Uhlen,M. and Flock,J.-I. (1985) *Nucleic Acids Res.*, **13**, 8913–8926 (X03341). Smith,E.L., Delange,R.J., Evans,W.H., Landon,M. and Markland,F.S. (1968) *J. Biol. Chem.*, **243**, 2184–2191 (PIR A00968; amino acid sequencing; mature protease sequence differs in having T103S, P129A, S158N, N161S and S212N).

BLS147 Hastrup,S., Branner,S., Norris,F., Petersen,S.B., Norskov-Lauridsen,L., Jensen,V.J. and Aaslyng,D. (1989) PCT Patent Appl. WO 8906279. Pub. July 13 1989. Appl. DK 8900002 filed Jan. 6, 1989 (Esperase™). Takami,H., Akiba,T. and Hoikoshi,K. (1990) *Appl. Microbiol. Biotechnol.*, **33**, 519–523 (amino acid sequencing of mature alkaline protease residues 1–20 from *Bacillus* sp. no. AH-101; this sequence differs from BLS147 in having N11S).

BABP92 van der Laan,J.C., Gerritse,G., Mulleners,L.J.S.M., van der Hoek,R.A.C. and Quax,W.J. (1991) *Appl. Environ. Microbiol.*, **57**, 901–909. (MaxacalTM). Hastrup,S., Branner,S., Norris,F., Petersen,S.B., Norskov-Lauridsen,L., Jensen,V.J. and Aaslyng,D. (1989) PCT Patent Appl. WO 8906279. Pub. 13 Jul 1989. Appl. DK 8900002 filed Jan. 6, 1989 (subtilisin 309, Savinase™, from *B.lentus* differs only in having N87S). Godette,D., Paech,C., Yang,S., Mielenz,J., Bystroff,C., Wilke,M. and Fletterick,R. (1991) Abstracts 5th Protein Society Symposium, June 22–26, Baltimore; abstract M8 (a high-alkaline protease from *B.lentus* differs in having N87S, S99D, S101R, S103A, V104I and G159S).

BDSM48 Rettenmaier,H., Kreimeyer,A., Perner,J. and Diessel,P. (1990) PCT Patent Appl. WO 90/04022. Publ. April 19, 1990. Appl. DE P3834550.1 filed Oct. 11, 1988.

BYSYAB Kaneko,R., Koyama,N., Tsai,Y-C., Juang,R.-Y., Yoda,K. and Yamasaki,M. (1989) *J. Bacteriol.*, **171**, 5232–5236 (M28537).

BSEPR Sloma,A., Ally,A., Ally,D. and Pero,J. (1988) *J. Bacteriol.*, **170**, 5557–5563 (M22407). Bruckner,R., Shoseyov,O. and Doi,R.H. (1990) *Mol. Gen. Genet.*, **221**, 486–490 (X53307).

BSBPF Sloma,A., Rufo,G.A.,Jr, Rudolph,C.F., Sullivan,B.J., Theriault,K.A. and Pero,J. (1990) *J. Bacteriol.*, **172**, 1470–1477 (M29035; corrected). Wu,X.-C., Nathoo,S., Pang,A.S.-H., Carne,T. and Wong,S.-L. (1990) *J. Biol. Chem.*, **265**, 6845–6850 (J05400; this sequence differs in having A169V and 586 less C-terminal residues due to a frameshift).

BSISP1 Koide,Y., Nakamura,A., Uozumi,T. and Beppu,T. (1986) *J. Bacteriol.*, **167**, 110–116 (M13760).

BSIA50 Strongin,A.Ya., Izotova,L.S., Abramov,Z.T., Gorodetsky,D.I., Ermakova,L.M., Baratova,L.A., Belyanova,L.P. and Stepanov,V.M. (1978) *J. Bacteriol.*, **133**, 1401–1411 (amino acid sequencing of mature protease residues 1–54; residues 3, 39, 40, 45, 46, 49 and 50 not determined).

BTFINI Chestukhina,G., Zagnitko,O.P., Revina,L.P., Klepikova,S. and Stepanov,V.M. (1985) *Biokhimiya*, **50**, 1724–1730 (amino acid sequencing of mature protease residues 1–14 from *B.thuringiensis* variety *israelensis*, and residues 1–16 and 223–243 from variety *finitimus*). Kunitate,A., Okamoto,M. and Ohmori,I. (1989) *Agric. Biol. Chem.*, **53**, 3251–3256 (amino acid sequencing of mature protease residues 6–20 from variety *kurstaki*, BTKURS).

BCESPR Chestukhina,G.G., Zagnitko,O.P., Revina,L.P., Klepikova,S. and Stepanov,V.M. (1985) *Biokhimiya*, **50**, 1724–1730 (amino acid sequencing of mature residues 1–16 and 223–243).

NDAPII Tsujibo,H., Miyamoto,K., Hasegawa,T. and Inamori,Y. (1990) *Agric. Biol. Chem.*, **54**, 2177–2179 (amino acid sequencing of mature residues 1–26).

TVTHER Meloun,B., Baudys,M., Kostka,V., Hausdorf,G., Frommel,C. and Hohne,W.E. (1985) *FEBS Lett.*, **183**, 195–200 (PIR A00973; amino acid sequencing of mature protease residues 1–279).

EFCYLA Segarra,R.A., Booth,M.C., Morales,D.A., Huycke,M.M. and Gilmore,M.S. (1991) *Infect. Immun.*, **59**, 1239–1246.

SEEPIP Schnell,N., Engelke,G., Augustin,J., Rosenstein,R., Götz,F. and Entian,K.-D. (1991), personal communication.

SPSCPA Chen,C.C. and Cleary,P.P. (1990) *J. Biol. Chem.*, **265**, 3161–3167 (J05229).

DNEBPR Kortt,A.A., Lilley,G.G. and Stewart,D.T. (1991) Abstracts 5th Protein Society Symposium, June 22–26, Baltimore, abstract S76.

LLSK11 Vos,P., Simons,G., Siezen,R.J. and De Vos,W.M. (1989) *J. Biol. Chem.*, **264**, 13579–13585 (J04962). Kok,J., Leenhouts,K.J., Haandrikman,A.J., Ledeboer,A.T. and Venema,G. (1988) *Appl. Environ. Microbiol.*, **54**, 231–238 (M24767; the sequence from strain Wg2 differs in 44 positions, including 18 differences in the protease domain, and a deletion of residues 1617–1676). Kiwaki,M., Ikemura,H., Shimizu-Kadota,M. and Hirashima,A. (1989) *Mol. Microbiol.*, **3**, 359–369 (X14130; the sequence from strain NCD0763 differs in 46 positions, including 22 in the protease domain, and a deletion of residues 1617–1676).

XCEXPR Liu,Y.-N., Tang,J.-L., Clarke,B.R., Dow,J.M. and Daniels,M.J. (1990) *Mol. Gen. Genet.*, **220**, 433–440.

SMEXSP Yanagida,N., Uozumi,T. and Beppu,T. (1986) *J. Bacteriol.*, **166**, 937–994 (M13469).

TAAQUA Terada,I., Kwon,S.-T., Miyata,Y., Matsuzawa,H. and Ohta,T. (1990) *J. Biol. Chem.*, **265**, 6576–6581 (J05414).

TRT41A McHale,R.H., Luthi,E., Saul,D.J., Ashby,M., Whitfield,J., Bergquist,P.L. and Clarke,N.H. (1990) Abstracts 5th Eur. Congr. Biotechn. Christiansen,C., Munck,L. and Villadsen,J. (eds), Munksgaard Int. Publishers, Copenhagen.

VAPROA Deane,S.M., Robb,F.T., Robb,S.M. and Woods,D.R. (1989) *Gene*, **76**, 281–288 (M25499).

SRESPD Lavrenova,G.I., Gul'nik,S.V., Kalugar,S.V., Borovikova,V.P., Revina,L.P. and Stepanov,V.M. (1984) *Biochemistry USSR*, **49**, 447–454 (amino acid sequencing of residues 1–23; residues 13, 18 and 19 not determined).

AVPRCA Maldener,I., Lockau,W., Cai,Y. and Wolk,C.P. (1991) *Mol. Gen. Genet.*, **225**, 113–120 (the published sequence has 28 uncertain residues near position 200–210 due to a frameshift reading error).

TAPROK Gunkel,F.A. and Gassen,H.G. (1989) *Eur. J. Biochem.*, **179**, 185–194 (X14688/X14689). Jany,K.D., Lederer,G. and Mayer,B. (1986) *Biol. Chem. Hoppe-Seyler*, **367**, 87 (PIR A24541; amino acid sequencing; mature protease differs in having S74SG, SILST204-208DSL and VNLL264-267FNL).

TAPROR Samal,B.B., Karan,B., Boone,T.C., Osslund,T.D., Chen,K.K. and Stabinsky,Y. (1990) *Mol. Microbiol.*, **4**, 1789–1792 (X56116).

TAPROT Samal,B.B., Karan,B., Boone,T.C., Chen,K.K., Rohde,M.F. and Stabinsky,Y. (1989) *Gene*, **85**, 329–333.

AOALPR Tatsumi,A., Ogawa,Y., Murakami,S., Ishida,Y., Murakami,K., Masaki,A., Kawabe,H., Arimura,H., Nakano,E. and Motai,H. (1989) *Mol. Gen. Genet.*, **219**, 33–38. Cheevadhanarah,S., Saunders,G., Renno,D.V., Holt,G. and Flegel,T. (1991) EMBL Data Library (X54726).

MPTHMY Gaucher,G.M. and Stevenson,K.J. (1976) *Methods Enzymol.*, **45**, 415–433 (amino acid sequencing of residues 1–28, and hexapeptide LSGTSM with active site serine).

ACALPR Isogai,T., Fukugawa,M., Kojo,H., Kohsaka,M., Aoki,H. and Imanaka,H. (1991) *Agric. Biol. Chem.*, **55**, 471–477. Stepanov,V.M., Rudenskaya,G.N., Vasil'eva,L.I., Krest'anova,I.N., Khodova,O.M. and Bartoshevitch,Y.E. (1986) *Int. J. Biochem.*, **18**, 369–375 (amino acid sequencing of residues 1–27; the mature protease differs in having H13[1]Q, R13[2]N and S13[6]A).

KLKEX1 Tanguy-Rougeau,C., Wesolowski-Louvel,M. and Fukuhara,H. (1988) *FEBS Lett.*, **234**, 464–470 (X07038).

SCKEX2 Mizuno,K., Nakamura,T., Ohshima,T., Tanaka,S. and Matsuo,H. (1988) *Biocnem. Biophys. Res. Commun.*, **156**, 246–254 (M24201).

Table I. Continued

SCPRB1	Moehle,C.M., Tizard,R., Lemmon,S.K., Smart,J. and Jones,E.W. (1987) *Mol. Cell. Biol.*, **7**, 4390–4399 (M18097).
YLXPR2	Davidow,L.S., O'Donnell,M.M., Kaczmarek,F.S., Pereira,D.A., DeZeeuw,J.R. and Franke,A.E. (1987) *J. Bacteriol.*, **169**, 4621–4629 (M17741). Matoba,S., Fukayama,J., Wing,R.A. and Ogrydziak,D.M. (1988) *Mol. Cell. Biol.*, **8**, 4904–4916 (M23353).
CEBLI4	Peters,K. and Rose,A. (1991) *The Worm Breeder's Gazette*, **11**, 28.
DMFUR1	Roebroek,A.J.M., Pauli,I.G.L., Zhang,Y. and van de Ven,W.J.M. (1991) *FEBS Lett.*, in press (X59384).
DMFUR2	Roebroek,A.J.M. *et al.*, in preparation.
CMCUCU	Kaneda,M., Ohmine,H., Yonezawa,H. and Tominaga,N. (1984) *J. Biochem.*, **95**, 825–829 (amino acid sequencing of octapeptide NIISGTSM with active site serine).
HSFURI	van den Ouweland,A.M.W., van Duijnhoven,H.L.P., Keizer,G.D., Dorssers,C.J. and van de Ven,W.J.M. (1990) *Nucl. Acids Res.*, **18**, 664 (X04329). van de Ven,W.J.M., personal communication (the sequence of mouse furin differs in 51 positions, including five in the catalytic domain: A15E, Y21F, S223F, A232V and N258[2]D). Misumi,Y., Sohda,M. and Ikehara,Y. (1990) *Nucl. Acids Res.*, **18**, 6719 (X55660; the sequence of rat furin differs in 49 positions, including three in the catalytic domain: A15E, Y21F, H24R).
HSIPC2	Smeekens,S.P. and Steiner,D.F. (1990) *J. Biol. Chem.*, **265**, 2997–3000 (J05252). Seidah,N.G., Gaspar,L., Mion,P., Marcinkiewicz,M., Mbikay,M. and Chretien,M. (1990) *DNA Cell Biol.*, **9**, 415–424 (the sequence of mouse pituitary PC2 protease differs in 23 positions, including seven in the protease domain: I9F, S42[2]Y, E45D, N76S, D133E, V134L, and G239[1]D).
MMPPC3	Smeekens,S.P., Avruch,A.S., LaMendola,J., Chan,S.J. and Steiner,D.F. (1991) *Proc. Natl Acad. Sci. USA*, **88**, 340–344 (M58507). Seidah,N.G., Gaspar,L., Mion,P., Marcinkiewicz,M., Mbikay,M. and Chretien,M. (1900) *DNA Cell Biol.*, **9**, 415–424 (M55668/M55669; partial sequence).
HSTPP	Tomkinson,B. and Jonsson,A.-K. (1991) *Biochemistry*, **30**, 168–174 (J05299).

protein-engineering strategies for each of the subtilases, aimed at modulating either stability, catalytic activity or substrate specificity.

Materials and methods

Comparison of atomic models

Atomic coordinates were obtained from the Brookhaven Protein Data Base for subtilisin BPN′ (PDB code 2SNI; McPhalen and James, 1988), subtilisin Carlsberg (1CSE; Bode *et al.*, 1987), thermitase (1TEC; Gros *et al.*, 1989) and proteinase K (2PRK; Betzel *et al.*, 1988a,b,c). Pairwise superposition of atomic coordinate models (Cα atoms only) was performed using the method of Kabsch (1976). Secondary structure elements and residue accessibility in the atomic models were analysed with the DSSP program (Kabsch and Sander, 1983).

Database search

In addition to literature searches, the EMBL (release 26.0), Swiss-Prot (release 17.0) and NBRF/PIR (release 27.0) databases were searched using the program FASTA (Pearson and Lipman, 1988) through the facilities of the CAOS/CAMM Center, Nijmegen, The Netherlands. Consensus sequence segments around the active site residues H64 and S221 were used for this purpose (see Figure 3).

Amino acid sequence alignment

Multiple sequence alignment was initially performed using the CLUSTAL program (Higgins and Sharp, 1988). Next, alignment improvements were made using the SALE program (J.Leunissen, unpublished results), taking into account the known alignment from superposition of three-dimensional structures (see Figure 2). The amino acid numbering used throughout this paper corresponds to that of mature subtilisin BPN′, our reference sequence. Residues in inserts relative to this reference sequence are numbered in square brackets; for instance, residues inserted between positions 12 and 13 are numbered 12[+1], 12[+2], etc.

Construction of phylogenetic trees

Distance matrices for all sequences were constructed using the minimum mutation distances (Fitch and Margoliash, 1967). Only residues of the catalytic domain (see Figure 3) were used in the calculations. Phylogenetic trees were constructed using the program FITCH (Felsenstein, 1990), according to the method of Fitch and Margoliash (1967), modified to disallow negative branch lengths (Prager and Wilson, 1978). The program employs local and global branch swapping to find the minimum length tree. It was run several times, using different (random) input orders, to overcome any dependency on the input order of the minimum tree found. All runs found the same minimum-length tree. The programs KITSCH (Felsenstein, 1990), UPGMA (Sneath and Sokal, 1973; J.Miller, personal communication) and NJTREE (Saitou and Nei, 1987; J.Miller, personal communication) all produced trees which reflected the general topology as calculated from FITCH, with only minor differences in the branching pattern. The tree was rerooted prior to printing with TREEMOVE (J.Leunissen, modified after DNAMOVE, Felsenstein, 1990) and was printed using DRAWGRAM (Felsenstein, 1990).

Results

Comparison of known three-dimensional structures

The Cα atoms of subtilisin BPN′, subtilisin Carlsberg, thermitase and proteinase K were superimposed to obtain a maximal overlap of the backbone structures (Figure 1A). Large parts of all four structures overlap very well. The results of detailed pairwise superpositions are summarized in Figure 2, in which the sequence alignment shown corresponds to structural equivalence of the Cα atoms. Topologically equivalent residues which have Cα-atom distances of <1.7 Å are defined as the core or 'structurally conserved regions', abbreviated SCRs (Greer, 1990).

In the comparison of all four structures together we can identify 194 structurally equivalent core residues in this way, assuming a minimum of three consecutive core residues (black bars in Figure 2). External helices, such as hD, hE, hG and the end of hF, which are slightly shifted or rotated relative to each other such that not all Cα atoms are within 1.7 Å, are also included in the core, since the helices are clearly structurally equivalent (Figure 1B). Table II shows the root-mean-square deviation values for the pairwise superposition of these 194 Cα atoms, together with the percentage amino acid sequence identity of the corresponding residues. A clear correlation is seen between these two parameters, which indicates that higher sequence identity corresponds to a closer overlap of main-chain atoms in the core (Chothia and Lesk, 1986).

When only the subtilisin BPN′, subtilisin Carlsberg and thermitase structures are superimposed the number of structurally equivalent Cα atoms increases to 232 (or ~85% of all Cα atoms), which we refer to as the 'extended core' (black and white bars in Figure 2). This distinction between core and extended core SCRs is of relevance for homology modelling, since the

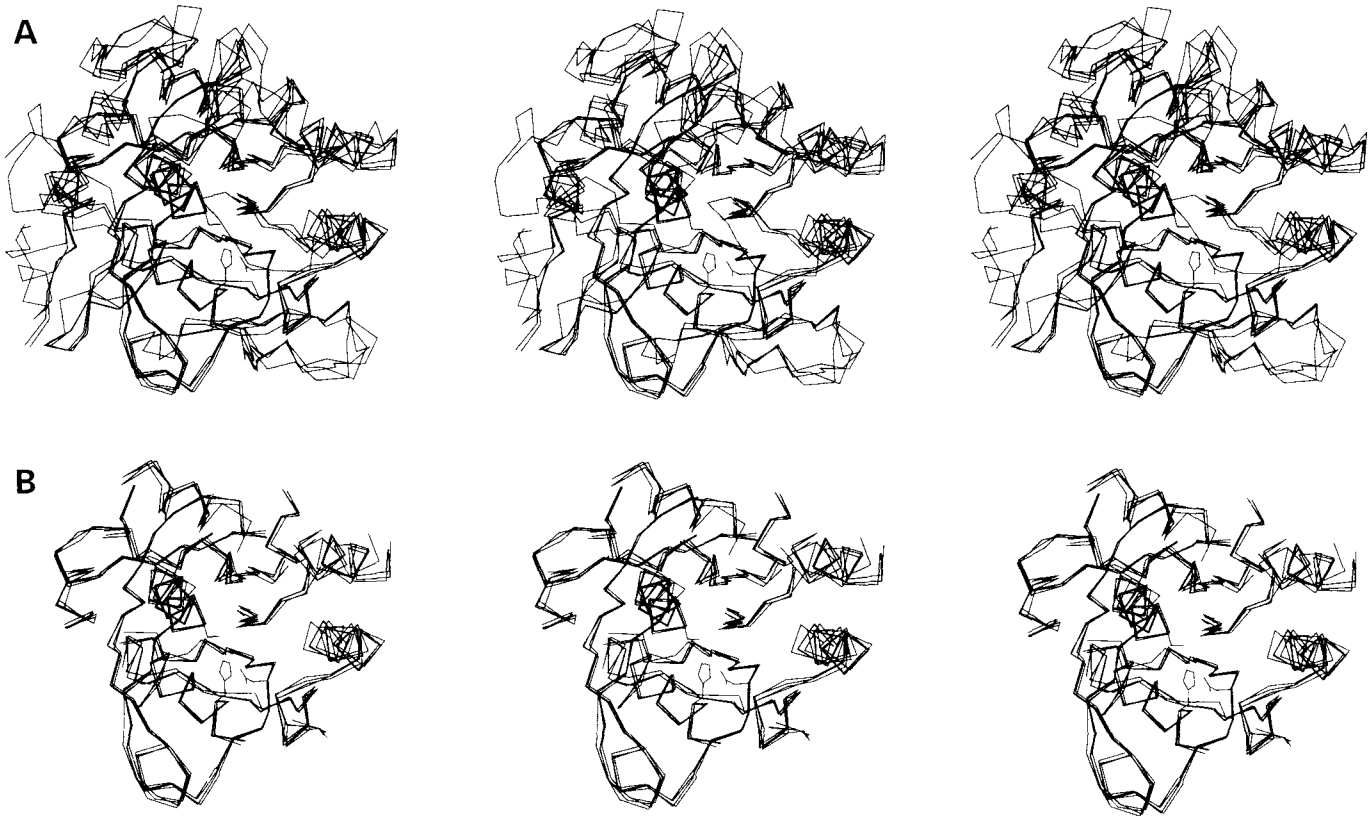


Fig. 1. Stereo diagram of the superposition of the Cα-atom backbones of subtilisin BPN′, subtilisin Carlsberg, thermitase and proteinase K. Only the side chains of catalytic triad residues Asp32, His64 and Ser221 are shown. (**A**) Complete backbones, and (**B**) core of 194 residues defined as the structurally conserved regions (SCRs).

family of subtilases can be subdivided into two main classes, I and II, as discussed below.

The common secondary structure elements in all four enzymes are also shown in Figure 2. We see that the core contains virtually all of the common α-helix and β-strand elements, including the active site residues D32, H64 and S221 which are located at the ends of β-strand e1 and the helices hC and hF respectively. Highly detailed comparisons of pairs of three-dimensional structures have been presented recently: subtilisins BPN′ and Carlsberg (Bode *et al.*, 1987; McPhalen and James, 1988), subtilisin BPN′ and thermitase (Teplyakov *et al.*, 1990), subtilisin Carlsberg and thermitase (Frommel and Sander, 1989) and thermitase and proteinase K (Betzel *et al.*, 1990).

The connections between core segments can differ considerably between members of the family, both in length and in amino acid sequence. We refer to these connections as 'variable regions' or VRs. The VRs nearly always correspond to connecting loops between helices and β-strands and generally lie on the external surface of the protein. The structural non-equivalence of these loops may result from amino acid additions or deletions, or they may result from loop flexibility or thermal motion (Gros *et al.*, 1990).

We assume that the structurally conserved core as defined above will usually be conserved in other homologous members of the subtilase family, as has been observed in other protein families (Chothia and Lesk, 1986). This structural framework can then be used for homology modelling of subtilases of known primary structure but unknown three-dimensional structure.

Identification of subtilase family members

An extensive search of scientific and patent literature and databases was performed to identify new subtilisin-like serine proteases, and the main results are summarized in Table I. Additional variants are described in the footnotes to the table. At present, >40 complete amino acid sequences of subtilisin-like proteinases (=endopeptidases) are known, and most are derived from the corresponding gene or cDNA sequences. Most recently, a human tripeptidyl peptidase was identified as a member of the subtilase family (Tomkinson and Jonsson, 1991). In addition, >10 other subtilases, for which only the N-terminal sequences are available, have been identified. The inclusion of the plant enzyme cucumisin is tentative, since only the sequence of an octapeptide with the putative active site serine is known.

Nearly all of the subtilases are synthesized as pre-pro-enzymes, subsequently translocated over a membrane, and finally activated by cleavage of the pro-segment. The known pre-pro-segments range from 27 to 280 residues. A detailed comparison of the pre-pro-sequences and the putative processing sites of these subtilases is in preparation (R.J.Siezen, unpublished results).

Table I shows that the (putative) mature enzymes range in size from 268 to 1775 residues. The catalytic domain or module is defined as the segment with sequence homology to subtilisins; it is always located at the N-terminal end, and ranges in size from 268 to 511 residues. The function of C-terminal extensions located after the catalytic domain is generally not clear; some may contain structurally distinct domains with other functions. In one instance (LLSK11), several residues located far from the homologous catalytic domain were also found to affect catalytic activity (Vos *et al.*, 1991). Occasionally the C-terminal extension contains sequence repeats, and in several cases a membrane anchor sequence can be identified at the extreme C-terminus (Mizuno *et al.*, 1988; Vos *et al.*, 1989; Chen and Cleary, 1990; van de Ven *et al.*, 1990). In other cases, C-terminal processing occurs and may be required for activation or transport across a second

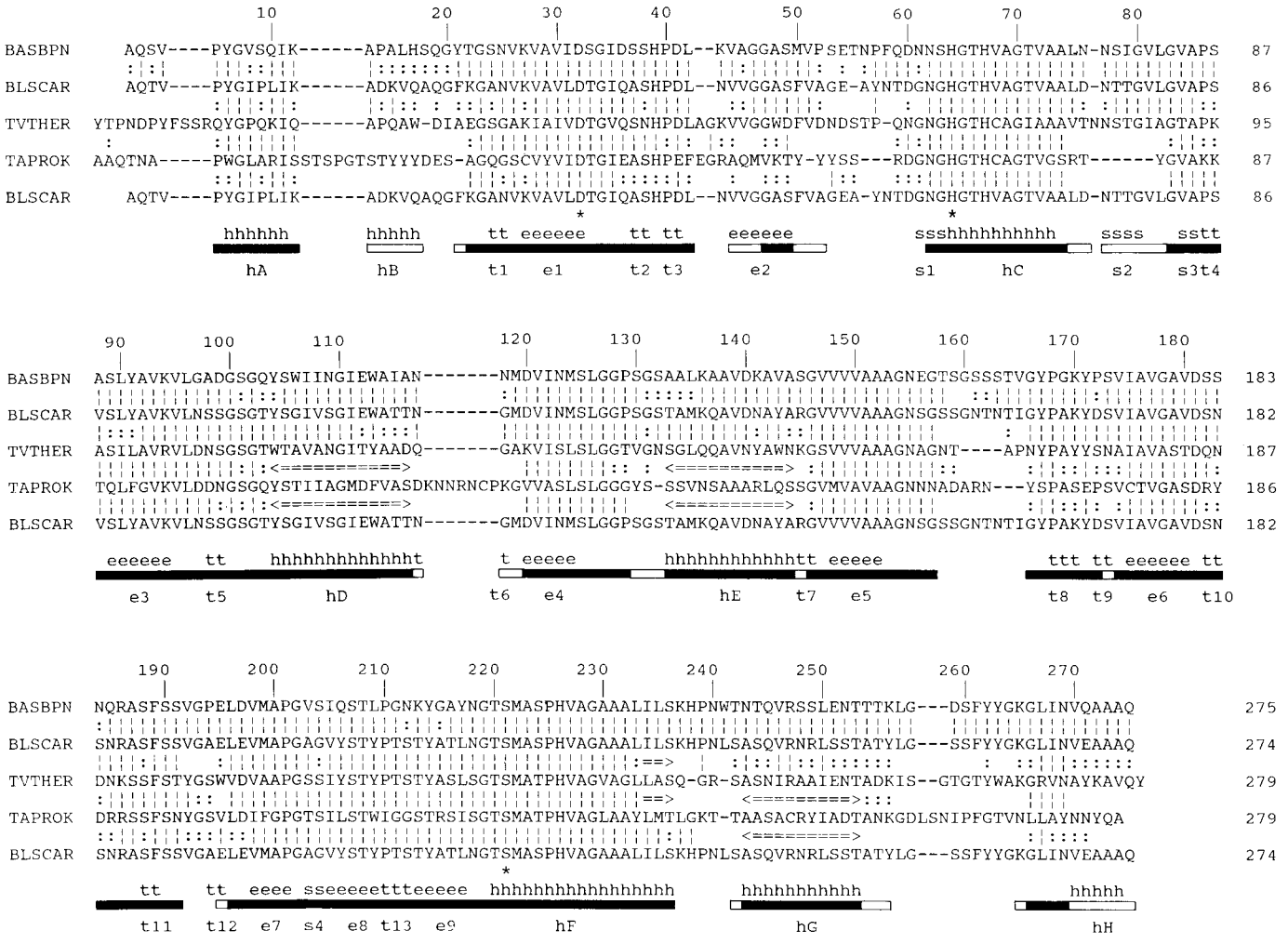


Fig. 2. Alignment of topologically equivalent amino acid residues of subtilisin BPN′ (BASBPN), subtilisin Carlsberg (BLSCAR), thermitase (TVTHER) and proteinase K (TAPROK) deduced from the three-dimensional superposition. Residue numbering for each sequence is shown at the right, and for the reference BASBPN sequence also along the top. *, Active site residues. Topologically equivalent residues are indicated as follows: (¦) Cα-atom distance <0.8 Å; (:) Cα-atom distance between 0.8 Å and 1.7 Å; (<===>) displaced helix for which most residues have distances of Cα-atoms of <0.8 Å, some >1.7 Å, and the rest between 0.8 and 1.7 Å. Filled bars indicate core segments (SCRs) with Cα-atom distances <1.7 Å in all four structures. Open bars indicate additional segments of the 'extended core', based only on BASPN, BLSCAR and TVTHER. Common secondary structure elements in all four X-ray structures are shown as: h = helix, e = extended β-sheet, s = bend and t = β-turn. Exceptions which are not found in TAPROK are hB, hH, s2 and t6.

membrane such as the outer membrane in Gram-negative bacteria (Yanagida *et al.*, 1986; Terada *et al.*, 1990). In many cases the precise length of the mature, active enzyme is not known due to the fact that the N-terminal and/or C-terminal processing site(s) have not yet been mapped.

Alignment of primary sequences

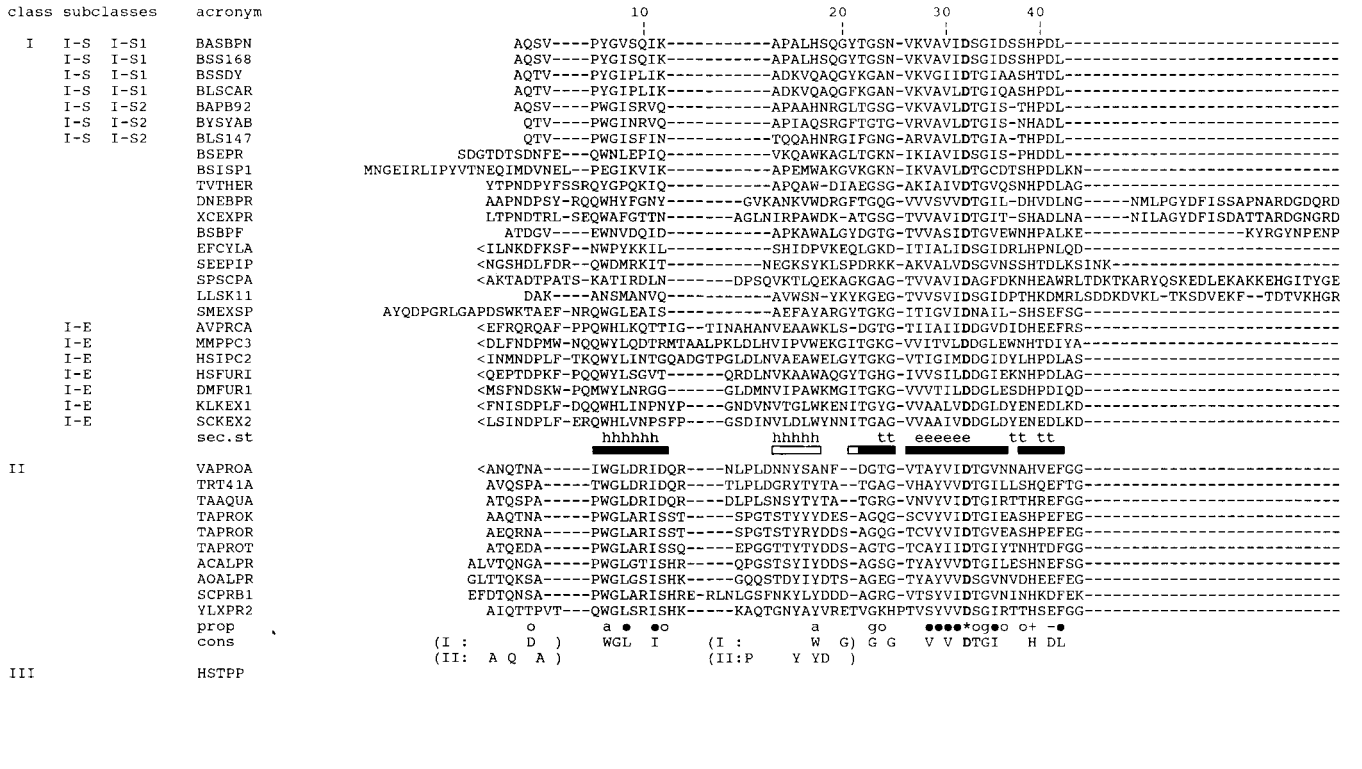
Figure 3 shows the amino acid sequence alignment, of the catalytic domains only, of a large number of subtilases to those of the four known structures. The core and extended core SCRs, as defined in Figure 2, are indicated by bars; the multiple sequence alignment indicates that residues 37, 98 and 99 should be omitted from these SCRs (*vide infra*). In every enzyme the sequence patterns that are characteristic of most of these SCRs are readily identified, and this facilitates the overall aligning and subsequent homology modelling. The alignments improve as the number of known sequences increases (at the current rate of one or two each month), since characteristic patterns, subclasses and allowed variations become more clear. However, the tripeptidyl peptidase sequence (HSTPP), shown at the bottom of Figure 3, is quite distinct from all other proteinase sequences: regions around the active site H64 and S221 are readily identified, but

Table II. Root-mean-square deviation values (Å) for the superposition of Cα-atoms (upper right) and percentage amino acid sequence identity (lower left) of the 194 core residues

	Subtilisin BPN′	Subtilisin Carlsberg	Thermitase	Proteinase K
Subtilisin BPN′	–	0.405	0.646	0.888
Subtilisin Carlsberg	75	–	0.621	0.940
Thermitase	52	56	–	0.949
Proteinase K	44	42	41	–

the homology at the N-terminus is so low that the SCR around active-site residue D32 cannot be identified, even though several Asp residues occur in this N-terminal region of HSTPP. Therefore, we shall not include it in the following analyses and comparisons.

From Figure 3 it is apparent that on the basis of sequence alignment the subtilases can be subdivided into class I and class II, shown above and below the core bars respectively. This distinction is based on characteristic sequence patterns and consensus residues (*vide infra*), both in SCRs and VRs, that are



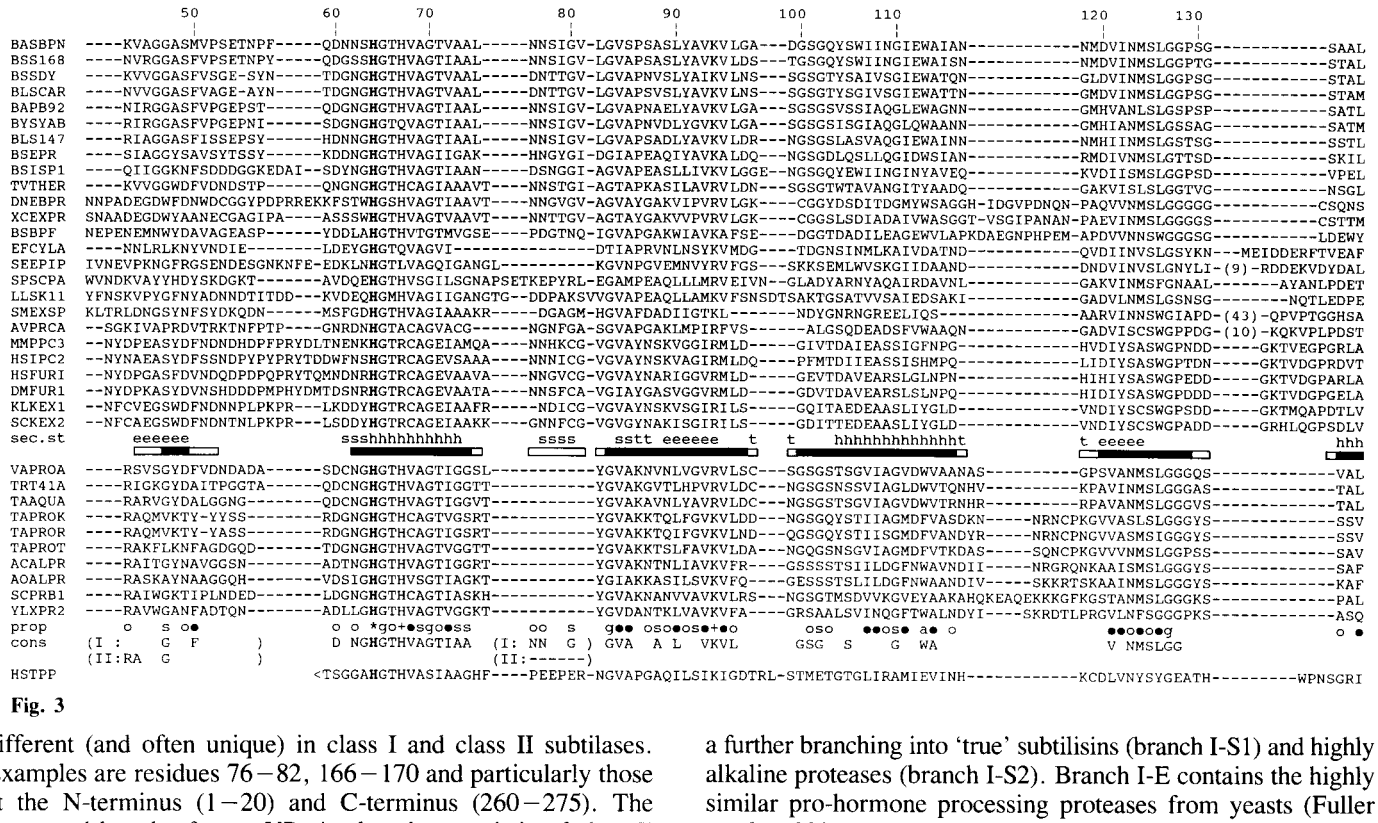


Fig. 3

different (and often unique) in class I and class II subtilases. Examples are residues 76−82, 166−170 and particularly those at the N-terminus (1−20) and C-terminus (260−275). The conserved length of most VRs is also characteristic of class II enzymes (*vide infra*).

These two main classes are particularly clear in a family tree or cladogram (Figure 4), a measure of the sequence homology between family members, constructed from the sequence alignment of the catalytic domains in Figure 3. Not only does the tree clearly depict the distinct class II subtilases, but it also demonstrates a few subtypes within the class I subtilases. The subtilisins from *Bacillus* strains are found in branch I-S, with a further branching into 'true' subtilisins (branch I-S1) and highly alkaline proteases (branch I-S2). Branch I-E contains the highly similar pro-hormone processing proteases from yeasts (Fuller *et al.*, 1989) and higher eukaryotes (van de Ven *et al.*, 1990). Other subclasses should become evident as new sequences become available.

Alignment of class II subtilases is fairly unambiguous due to their high degree of sequence homology, even in most VRs, and the low incidence of insertions/deletions relative to proteinase K. The C-terminal residues from position 240 onwards are the least homologous, but the alignment is still clear. Only three of these class II enzymes have a C-terminal extension beyond the

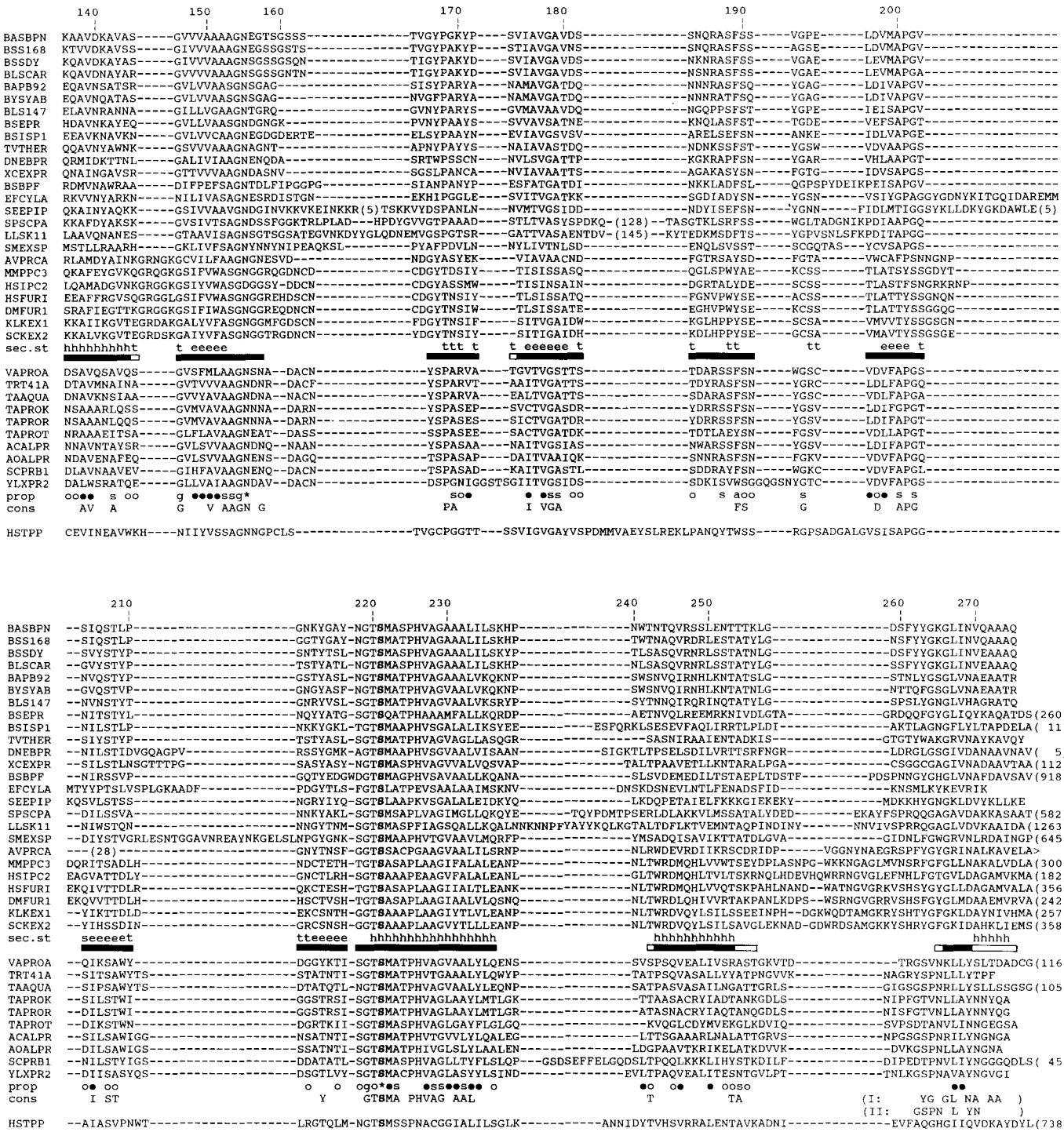


Fig. 3. Alignment of amino acid sequences of catalytic domains of subtilases. Enzyme acronyms are given in Table I. Residue numbering at the top corresponds to that of mature subtilisin BPN′ (BASBPN). The class I and class II subtilases (see text) are separated by filled bars (SCRs or core segments) and open bars (extended core; applicable only to class I), and the secondary structure assignments, as defined in Figure 2. Below the aligned sequences the consensus physical properties (prop) and consensus amino acids (cons) at each position are shown. A consensus physical property is defined as >75% of residues belonging to one of the following groups: nonpolar, A, V, I, L, M, C, F, W, Y (●); polar, S, T, N, Q, D, E, R, K, H (○); small, G, A, S, C, V (s); glycine, G (g); aromatic, F, Y, W (a); negative charge, D, E (−); positive charge, R, K, H (+); active site residues D32, H64, S221, shown in bold (*). A consensus amino acid occurs in >50% of all sequences; consensus residues in brackets apply either to class I or class II subtilases only, as indicated. The number of additional residues in large inserts in the catalytic domain, and in C-terminal extensions, are shown in brackets. Each sequence begins at the mature N-terminus, if known; an unknown number of residues is indicated as (<) at the N-terminus and as (>) at the C-terminus. Partial alignment of the sequence of the human tripeptidyl peptidase (HSTPP) from residue 59 onwards is shown on the bottom line, as a separate class III subtilase.

catalytic domain. Alignment of N-terminal parts of incompletely sequenced serine proteases (Figure 5) indicates that NDAPII, SRESPD and MPTHMY are also class II subtilases.

Alignment of class I subtilases is straightforward for most of the SCRs. However, in some of the most highly diverged sequences there are regions with very weak sequence homology, even in the core, which results in alignments that are not unambiguous. In those cases, alternative alignments to those in

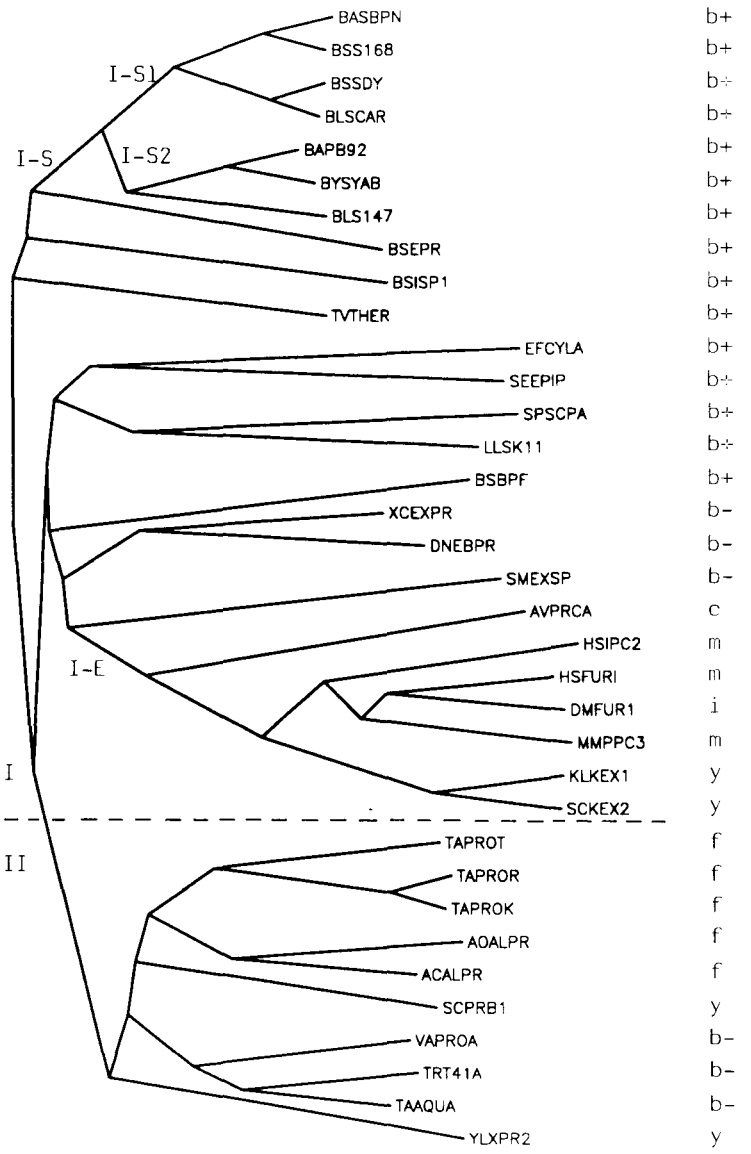


Fig. 4. A family tree or cladogram, schematically depicting the homology between catalytic domains of subtilases. Branch lengths are in inverse proportion to the degree of sequence similarity. The two main classes I and II and some subclasses are indicated. (b+) Gram-positive bacterium, (b−) Gram-negative bacterium, (f) fungus, (y) yeast, (i) insect, (m) mammal.

Figure 3 may need to be considered. These regions are found on the surface of the molecule and contain numerous solvent-exposed residues, allowing for greater side-chain variation. Examples are (i) the exposed regions 43−58 and 182−218 which contain structurally conserved β-strands and turns, and (ii) the exposed amphipathic helices hD (residues 104−116), hE (133−144) and hG (243−253). In the latter case, the sequence alignment of amphipathic helices is also based on the requirement that at certain positions non-polar side chains are conserved that point into the interior of the molecule, while polar residues face outward.

Based on the N-terminal sequence only, six other bacterial proteases can be classified as class I (Figure 5). Four of these, including the only archaeal subtilase known (ARB172), appear to be closely related to thermitase, and should cluster in a new thermitase branch (I-T) of the cladogram.

Appendix A summarizes all naturally occurring amino acids at topologically equivalent positions, together with various other structural and functional characteristics of each residue. Note that Appendix A is compiled only from the sequences in Figure 3.

Structurally conserved regions (SCRs)

The 194 residues that constitute the SCR core, as defined from the four known structures in Figure 2, are nearly all present in the other subtilases. One minor exception is the absence of residue 37 in several class I family members. Another is the absence of residues 73−74 in EFCYLA, shortening helix hC; the connection of residue 72 to 82 is feasible since they are nearly adjacent in subtilisin (Figure 4 in McPhalen and James, 1988). A major exception is the absence of various residues in the range 96−100 in 12 different subtilases, including all of those in class I-E. Structurally, this is quite plausible since it would eliminate part of an external loop whose ends are close together. Catalytically, major changes can be expected since this loop is involved in substrate binding (*vide infra*). These considerations lead us to redefine the SCR core for all subtilases to only 191

```
class   acronym                  10                  20          30
                                 |                   |           |
I       BASBPN       AQSV----PYGVSQIK-------APALHSQGYTGSNVKVAVIDSGID
        BDSM48         QTV----PCGIPYIY-------SRVVHRQGYFGNGV
        BSIA50       NVxEL----PEGIQVIK-------APQLWAQGFKGSDIKIAVLDTGID

        TVTHER     YTPNDPYFSSRQYGPQKIQ-------APQAW-DIAEGSGAKIAIVDTGVQ
        ARB172     ATPNDPQY-GQQYAPQQVN-------xEAAWDVTYGDPGVTIS
        BTKURS     WTPNDPYF-NNQYGLGKIQ-------AP
        BTFINI     WTPNDPYF-KNQYGLQKV
        BCESPR     WTPNDPYY-KNQYGLQKL

II      TAPROK     AAQTNA-----PWGLARISST-SPGTSTYYYDES-AGQGSCVYVIDTGIE
        NDAPII           V-----PWGLDRIDQEDLPLDGSYTT-ES-DGS
        SRESPD     ATQTNP----PSWGLDxIDQAxxPLEG
        MPTHMY   ALVTQsNA----PsWGLGRIsNR-QAGIRDYHY
```

Fig. 5. N-terminal sequence alignment of incompletely sequenced mature subtilases. Sequences in bold (BASBPN, TVTHER, TAPROK) are reference sequences of known three-dimensional structures. x, Residue not identified; lower case signifies an uncertain assignment.

residues, eliminating three residues that are absent in more than one sequence, viz. 37, 98 and 99. This core is shown schematically in Figure 6 as the set of conserved α-helices and β-sheet strands.

Highly conserved core residues are listed in Table III. Only 11 positions are totally conserved; they are found in the active site (D32, H64, S221), substrate-binding region (S125, G127, G154, G219, T220) and internal helices (G65, G70, P225). Another seven residues are varied only once or twice, including the oxyanion hole residue N155 (Table III). Nine of these 18 most conserved residues are glycines, many of which have main-chain torsion angles not allowed for amino acid residues with side chains. In addition, only two different amino acid residues are found at positions 68, 94, 123, 152, 193, 207, 223 and 229, while three different residues occur at positions 69, 153, 168, 177, 179, 187, 201, 205, 213 and 227. In general, the residues of the two internal helices hC (64–72) and hF (220–236) are the most highly conserved in all subtilases. These conserved sequences are most suitable for database searches for new family members (see Materials and methods).

When only class I subtilases are considered we find no additional conserved residues. On the other hand, the class II subtilases are more highly conserved and contain a total of 58 invariant residues. Protein engineering of these highly conserved SCR residues may lead to major structural alterations of the framework core and concomitant loss of stability and activity.

At numerous other positions in the SCRs there is a distinct conservation of physical properties of side chains, such as size, polar/nonpolar, aromatic (Figure 3); charge conservation is less frequent (see also below). For instance, this applies to the amphipathic helices hA, hB, hD, hE and hG, in which a tremendous sequence variability is allowed as long as the amphiphilic character remains intact. Engineering of these variable SCR residues is a better choice, and one should take account, or make use of these types of side-chain requirements.

Active site

All sequences contain the essential catalytic triad residues D32, H64 and S221. Moreover, the most highly conserved amino acid sequences, as shown in the consensus sequence at the bottom of Figure 3, occur around these active site residues, namely segments 28–35, 62–74 and 218–233. In addition, all sequences but one contain N155 (in a conserved segment 152–155), that helps to stabilize the oxyanion generated in the tetrahedral transition state (Carter and Wells, 1990). The only

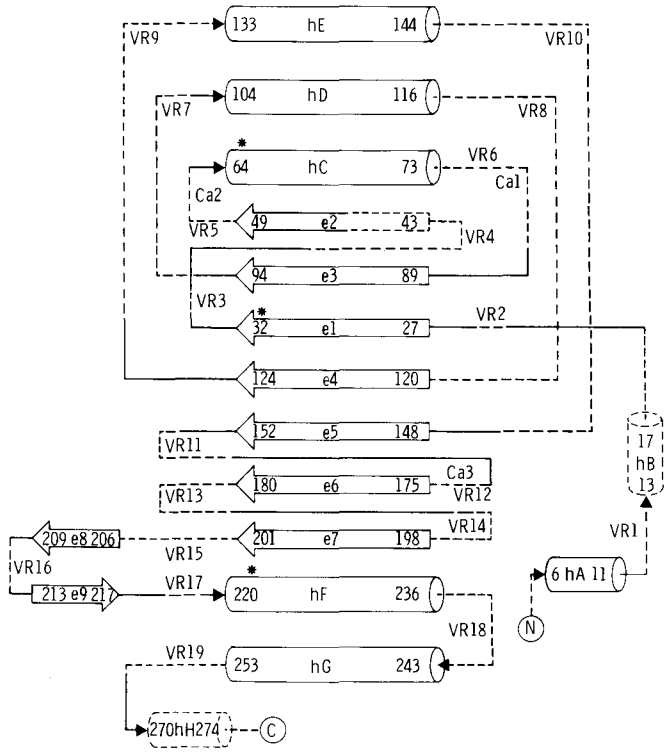


Fig. 6. A schematic representation of the secondary structural topology of subtilases, with α-helices shown as cylinders and β-sheet strands as arrows. Solid lines indicate the conserved regions (SCRs) in all subtilases, and dashed lines the variable regions (VRs). Approximate locations of the Ca1-, Ca2- and Ca3-binding sites are indicated, as are the catalytic triad residues (*).

accepted substitution here is N155D, as is found in HSIPC2, although the effect of this substitution on the catalytic efficiency of this putative protease has yet to be determined (Smeekens and Steiner, 1990).

Variable regions (VRs)

Table IV lists the positions and lengths of VRs in the subtilase family. Essentially the VRs comprise all the connections between conserved elements of secondary structure, as shown schematically in Figure 6. The VRs are generally located in loops on the surface of the molecule, allowing for variation in length and amino acid sequence. These VRs range in length from a

Table III. Highly conserved residues in subtilases

Position	Conserved residue	Alternative residue
23	G	R (1x)
32	D	
34	G	A (1x)
39	H	N (2x)
64	H	
65	G	
66	T	M,S (1x)
70	G	
83	G	T (1x)
125	S	
127	G	
146	G	D,N (1x)
154	G	
155	N	D (1x)
219	G	
220	T	
221	S	
225	P	

Table IV. Positions and lengths of the variable regions (VR) in subtilases

VR	Position	Residue length (x)		
		min.	max.	longest
N	x-5	3	≥22	BSISP1
1	12-x-22	8	20	MMPPC3, HSIPC2
2	25-x-26	0	1	YLXPR2
3	36-x-38	0	1	various
4	42-x-47	4	37	SPSCPA
5	49-x-62	8	17	various
6	74-x-83	−1	13	SPSCPA
7	97-x-102	−2	5	LLSK11
8	116-x-120	0	15	SCPRB1
9	129-x-133	2	52	SMEXSP
10	144-x-146	1	6	various
11	157-x-166	4	24	SEEPIP
12	172-x-174	0	5	YLXPR2
13	182-x-183	0	151	LLSK11
14	191-x-196	4	10	YLXPR2
15	203-x-204	0	22	SEEPIP
16	210-x-211	0	22	SMEXSP
17	217-x-218	0	1	BSBPF
18	236-x-243	4	21	LLSK11
19	253-x-266	12	28	HSIPC2
C	269-x	2	1270	LLSK11

single residue insertion or deletion (relative to our reference subtilisin BPN'), to a maximal deletion of nine residues and an insertion of 151 residues. Sequence identity or homology is also apparent in some VRs, particularly within class I, class I-E or class II subtilases.

Protein engineering of the VRs should present few limitations, and allows for variations of several properties of the enzyme with retention of the structural framework of the core. Some suggestions could be (i) to aim to increase thermal stability by modifying or shortening certain VRs to remove known autoproteolysis sites (e.g. VR11; Kim *et al.*, 1990), (ii) to introduce, alter or remove loops involved in binding of Ca^{2+} ions (e.g. VR5, VR6), and (iii) to alter loop sequence and/or length to affect catalytic activity and/or specificity (e.g. VR7). A further description of some of these examples follows below.

Homology modelling strategy

Model building of the three-dimensional structure of a protein from several homologous proteins of known structure is done in several steps; for details see, for example, Sutcliffe *et al.* (1987a, 1987b) and Greer (1981, 1990). In the first step the three-dimensional framework, or SCRs, for the family is defined. This provides the basis for modelling the protein backbone. In subsequent steps the VRs and side chains are modelled and energy minimized.

In the present case of the subtilases we have already defined the SCRs above. The following step could be to calculate an average framework (Blundell *et al.*, 1990) from the main-chain atom coordinates in SCRs of the four known three-dimensional structures. Alternatively, one could use the SCR main-chain coordinates from any of the individual known structures. In the latter case, the most related or homologous structure is preferably used, which can be found by inspection of the family tree (Figure 4).

Modelling of all class II subtilases should be based on the structure of proteinase K (TAPROK). Besides the SCRs, many of the VRs of TAPROK can be included in the model, since the VR lengths are often identical and the sequences reasonably homologous (e.g. VR4, VR6, VR9, VR10, VR11). In other VRs modelling of short insertions or deletions is required; the largest deletion is five residues (VR8) and the largest insertion 13 residues (VR18). The latter should pose problems, because large insertions are most difficult, if not impossible, to reliably predict for obvious reasons. In general, it should be possible to construct reasonably reliable models of all class II subtilases.

Modelling of class I subtilases can be based on the extended core of thermitase or either of the subtilisins BPN' and Carlsberg. In many cases the appropriate VRs from the known structure can be included in the model, particularly for most of the *Bacillus* proteases which have very few insertions or deletions (with the exception of BSBPF). On the other hand, in the other class I subtilases many VRs may prove extremely difficult to model, since very large insertions are observed in most of them. The most extreme example is LLSK11 which includes insertions of 30 (VR4), 14 (VR11), 151 (VR13) and 15 residues (VR18) compared with subtilisin, almost doubling the size of the catalytic domain.

Unique insertions and/or sequences in VRs are found in all class I-E subtilases, and these are quite conserved (e.g. VR4, VR5, VR9, VR10, VR11, VR14, VR19). It would be extremely helpful for modelling purposes to determine the three-dimensional structure of one of these unique class I-E subtilases, particularly since there is such a current interest in the biological function of these prohormone convertases.

Protein-engineering strategy

An extensive literature on engineering of subtilisin BPN' has appeared in recent years, partly summarized by Wells and Estell (1988). This engineering information, in combination with the sequence alignments and homology modelling described above, provides the basis for developing a protein-engineering strategy aimed at altering various catalytic and structural properties of any of the subtilases. Some examples will now be given.

Substrate specificity and catalysis

Residues that are in contact with substrate or inhibitor in thermitase, proteinase K, and subtilisins BPN' and Carlsberg have been identified from crystal structures and modelling of enzyme–inhibitor complexes (Hirono *et al.*, 1984; McPhalen *et al.*, 1985; Betzel *et al.*, 1988a; McPhalen and James, 1988;

Table V. Enzyme–substrate interactions in subtilases

Substrate residue	Subtilase residue	
	Conserved	Variable
P6		104, 128, 129, 130
P5	G127	102, 103, 104, 128
P4	G127	96, 100, 101, 102, 103, 104, 107, 126, 128
P3	G127	100, 101, 126
P2	H64, S125	33, 96, 100
P1	H64, S125, G127, G154, N155, G219, T220, S221	126, 152, 156, 166, 169
P1′	H64, N155, S221	217, 218
P2′	N155	189, 218, 219
P3′	H64	62, 209, 217

Substrate residue nomenclature is according to Schechter and Berger (1967), where the scissile peptide bond is between the P1 and P1′ substrate residues.
This table is based on known and modelled enzyme interactions with the following polypeptide segments:

Inhibitors:	-P6 -P5 -P4 -P3 -P2 -P1 -P1'-P2'-P3'-
eglin-c	- G - S - P - V - T - L - D - L - R -
chymotrypsin inhibitor-2	- G - T - I - V - T - M - E - Y - R -
Strept.subtilisin inhibitor	- D - V - M - C - P - M - V - Y - D -
Substrates:	F - A - A - Y - L - L
	suc- A - A - P - F -pNa

Gros *et al.*, 1989a,b), enzyme–substrate complexes (Robertus *et al.*, 1972; Wells *et al.*, 1987a) and from protein engineering of subtilisin BPN′ (Estell *et al.*, 1986; Russell and Fersht, 1987; Russell *et al.*, 1987; Wells *et al.*, 1987a,b). The enzyme residues known to interact with substrate or inhibitor residues P6 to P3′ are summarized in Table V; we have divided these enzyme residues into those invariant and those variable in all known subtilase sequences (see Figure 3). Main-chain and/or side-chain interactions between enzyme and substrate (or inhibitor) occur; details of possible hydrogen bonding, hydrophobic and electrostatic interactions can be found in the references above.

In general, the binding site can be described as a surface channel or crevice capable of accommodating at least six amino acid residues (P4–P2′) of a polypeptide substrate (or inhibitor). The N-terminal or P4–P1 specificity side of the substrate lines up between the extended enzyme backbone segments 100–103 and 125–128, forming the central strand of a three-stranded antiparallel β-sheet (McPhalen and James, 1988). The C-terminal or leaving portion P1′–P3′ of the substrate appears to be held less tightly as it runs along the enzyme backbone segment 217–219.

Substrate binding is predominantly determined by the binding of the P1 and P4 residues in two pockets or clefts on either side of the backbone strand 125–128. The two sides of the P1 cleft are formed by the backbone segments 125–128 and 152–155, while the segment 166–169 forms the bottom of the cleft. The P4 pocket, between the strands 101–104 and 126–128, is lined with hydrophobic side chains, i.e. residues 96, 104, 107 and 126 (Wells and Estell, 1988). In subtilisins and thermitase both pockets are large and hydrophobic, which explains the broad specificity of these enzymes with a preference for aromatic or large nonpolar P1 and P4 substrate residues. We propose that variations in the substrate specificity of naturally occurring subtilases should be due to (and can be modified by) modulation of the variable residues in Table V, and in the first place those residues whose side chains interact with P1 and P4 substrate residues.

Engineering studies of subtilisin BPN′ have demonstrated that P1 specificity can be dramatically modulated by substitutions of G166 at the bottom of the P1 cleft (Estell *et al.*, 1986; Wells *et al.*, 1987a,b). In addition, charged substitutions at positions 166 or 156 (at the P1 cleft entrance) shift the specificity to oppositely charged P1 residues (Wells *et al.*, 1987b). When we examine the position 166 in wild-type subtilases, we see that only YLXPR2 and SEEPIP have a negatively charged side chain (Asp); in addition they both contain a negatively charged Asp residue at position 156. This suggests that both enzymes should have a P1 specificity for positively charged residues, as was found for YLXPR2 (Ogrydziak, 1988). The only positively charged 166 residue found is a His in EFCYLA, suggesting a preference for negative P1 residues at low pH. In all other subtilases we find that at position 166 rather small, uncharged side chains (G, A, S, T, N) predominate, while at position 156 a side chain with negative charge or a hydroxyl group (D, E, S, T) is most commonly observed. Positive charges do not occur at position 156. Proteinase K and several related class II subtilases have a Tyr at position 166 (Figure 3), but since it is rotated away, the P1 cleft remains wide and hydrophobic (Betzel *et al.*, 1988a,b,c). This explains why proteinase K also has a broad P1 specificity with a slight preference for aromatic and hydrophobic residues.

In the P4 pocket, we find that hydrophobic side chains are generally conserved in wild-type subtilases, i.e. 96(L,F), 107 (I,V) and 126(L,W). Residue 104 at the pocket entrance varies considerably, from hydrophobic (Y,A,L) to hydrophilic (T,S,N,D), and this exposed residue is known to exhibit conformational flexibility in mutant subtilisins (McPhalen and James, 1988).

Additional enzyme–substrate interactions to those listed in Table V are to be expected in cases where side chains of enzyme and/or substrate differ from those in the known three-dimensional structures. For instance, since Gly residues are conserved at positions 100, 102 and 128 in all known three-dimensional structures, the possible contribution of other side chains at these positions to substrate binding is unknown. Furthermore, it has been argued that charged residues in lactococcal proteinases at positions 101 and 129, facing outward on opposite sides of the binding crevice, determine the specificity for substrates with charged P3 or P4 residues (Exterkate *et al.*, 1991; Vos *et al.*, 1991).

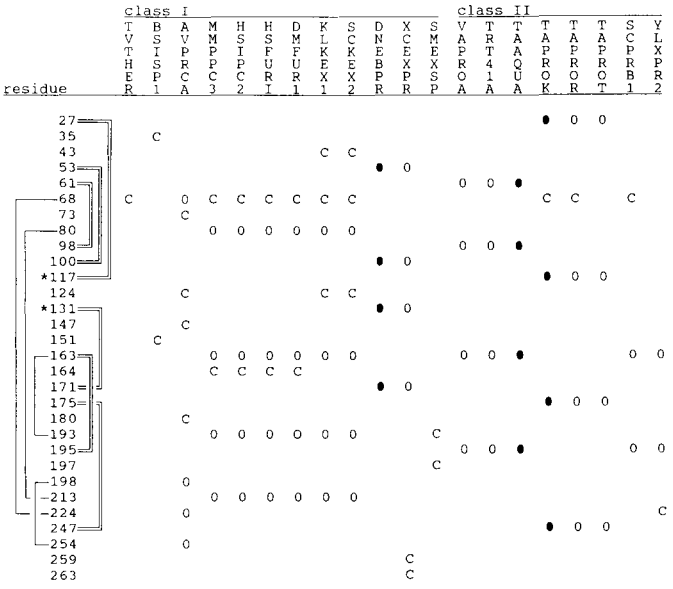
The high specificity of class I-E subtilases for paired basic residues, i.e. Arg/Lys at P2 and Arg at P1, may be facilitated by a high density of negative charge at the substrate-binding face (van de Ven *et al.*, 1990). This modelling study predicts that the (semi-) conserved residues D33, D61, D97, D104, E107, D130, D131, D161, D165 and D209 in class I-E subtilases are all in or near the substrate-binding region, and specific electrostatic interactions may occur with the positively charged P2 (e.g. with D33, D97) and P1 residues (e.g. with D165).

Finally, substitutions at position 31, next to the active-site residue D32, are known to markedly affect catalytic activity (Takagi *et al.*, 1988). Branched chain hydrophobic residues were found to be essential for activity, in the order L>I>V. Indeed, the only exception to this rule in wild-type subtilases is M31 in HSIPC2.

Oxidative stability

Chemical oxidation can be a significant source of enzyme inactivation, particularly for enzymes which function extracellularly. Oxidative stability can be improved by replacement of oxidatively sensitive residues, mainly cysteine and methionine, particularly those near the active site.

Table VI. Cysteine residues (C) and disulphides (●, known; ○, predicted) in subtilases



*Residues 117[+ 11] and 131[+ 1].
Filled circles and double connecting lines represent known S-S bonds in proteinase K (TAPROK), aqualysin (TAAQUA) and *Dichelobacter* protease (DNEBPR).
Open circles and single connecting lines represent predicted S-S bonds.

In subtilisin BPN′ the residues M124 and M222, both near the active site, are susceptible to oxidation by hydrogen peroxide; the bulky sulphoxide derivative near the catalytic site leads to a reduction of enzyme activity (Estell *et al.*, 1985). Protein-engineering studies showed that the best alternatives for M222 are the non-oxidizable residues Ala, Ser and Leu; this led to mutant subtilisins with retention of 12–53% activity and complete resistance to oxidation (Estell *et al.*, 1985). Such a mutant M222A (and G195E) subtilisin called Durazyme^{R} is presently used in a commercial detergent (Riisgaard, 1990).

Figure 3 shows that M222 is most common in wild-type subtilases. Natural variants are L222 (in EFCYLA), Q222 (in BSEPR), S222 (in AVPRCA) and indeed most frequently A222, found exclusively in class I-E subtilases. Interestingly, in the latter case A124 is the alternative for the other susceptible residue M124. Other natural variants of M124 are Leu, Phe, Val, Asn and Cys. Although C124 appears to be less suitable, since it is also oxidizable, all other natural variants should have improved oxidative stability.

Thermal stability: disulphides and cysteine residues

Disulphide bridges contribute to the overall stability of a protein, and the introduction of new S-S bonds can enhance the thermal stability, as demonstrated in, for example, phage T4 lysozyme (Masazumi *et al.*, 1989). Table VI shows all naturally occurring Cys residues in the catalytic domain of subtilases. The six disulphide bonds identified to date, as indicated by the double lines and filled circles in Table VI, are 27–117[+11] and 175–247 in proteinase K (TAPROK; Betzel *et al.*, 1988a,b,c), 61–98 and 163–195 in aqualysin (TAAQUA; Kwon *et al.*, 1988), both highly thermostable enzymes, and 53–100 and 131[+1]–171 in *Dichelobacter* protease (DNEBPR; A.Kortt *et al.*, personal communication).

Based on sequence homology (Figure 3), we predict that (i) the two disulphides of TAPROK also occur in the fungal enzymes TAPROR and TAPROT, (ii) the two disulphides of TAAQUA also occur in the bacterial enzymes VAPROA and TRT41A, (iii) the 163–195 S-S bond connecting the loops VR10 and VR13 should also occur in the yeast enzymes SCPRB1 and YLXPR2, and (iv) the two disulphides of DNEBPR also occur in XCEXPR at equivalent positions (Table VI).

Based on the known three-dimensional structures (Figure 1), together with the sequence alignment, we predict that four other natural S-S bonds may occur, as indicated by single connecting lines in Table VI. A nearly equivalent 163–193 (or 164–193) S-S bond to the 163–195 bond in aqualysin may occur in the class I-E subtilases. The uncertainty lies both in the fact that the C163 and C164 are almost adjacent, and in the fact that, although the VR10 and VR13 loops are close together, they differ somewhat in length compared with aqualysin. Another disulphide bond 80–213 may occur in all class I-E subtilases, as the residues C80 and C213 should also be close together in adjacent loops. Finally, an exceptional case is the cyanobacterial protease (AVPRCA), since five of its eight Cys residues are unique in their position compared with the other subtilases; the Cys residues predicted to be close enough together to form S-S bonds in AVPRCA are C68–C224 and C198–C254; the 68–224 disulphide would be buried directly under the active site.

All the remaining Cys residues in Table VI should not form S-S bonds, because they are either single (e.g. C68 in TVTHER, TAPROK, TAPROR and SCPRB1) or predicted to be too far removed from another Cys residue. The residues C68, C124 and C224 are internal and poorly accessible (Appendix A).

Cys residues are extremely rare in subtilases from Gram-positive bacteria and disulphides do not occur: only one Cys is found in TVTHER and two Cys residues located quite far apart topologically in BSISP1, an intracellular subtilase. This is a general feature of extracellular enzymes in Gram-positive bacteria. All other enzymes, with the exception of AOALPR and ACALPR, contain two to eight Cys residues which can presumably form a maximum of two disulphide bonds (Table VI).

Numerous attempts have been made, generally with little success, to increase the thermostability of subtilisins by the introduction of S-S bonds. Mutants with single disulphide bonds 22–87, 24–87, 25–232, 29–119, 36–210, 41–80 or 148–243 did not substantially improve stability (Katz and Kossiakoff, 1986, 1990; Wells and Powers, 1986; Pantoliano *et al.*, 1987; Mitchinson and Wells, 1989). However, none of these engineered disulphides corresponds to any of the (predicted) naturally occurring ones (Table VI). The only successful attempt was the introduction of the 61–98 S-S bond of aqualysin into subtilisin, thereby improving the thermal stability by 4.5°C (Takagi *et al.*, 1990). In the same study an attempt to introduce a 161–195 disulphide (rather than the natural 163–195 bond in aqualysin) was not successful; the Cys residues did not form an S-S bond.

A computer prediction of the energetically and stereochemically most acceptable sites for introduction of disulphides in subtilisin, under the assumption that no conformational changes occur, produced numerous possibilities (Hazes and Dijkstra, 1988). Only one of these predictions, 163–193, may correspond to a naturally occurring disulphide, as argued above.

We anticipate that introduction of one or more of the six known naturally occurring S-S bonds, summarized in Table VI, may be more successful in improving thermal stability.

Thermal stability: Ca^{2+}-binding sites

Binding of Ca^{2+} ions at specific sites, often in external loops, increases the stability of many proteases by reducing the flexibility

of the molecule and hence the denaturation and/or autolysis rate. Three Ca^{2+}-binding sites have been identified in the four known three-dimensional structures (McPhalen and James, 1988; Briedigkeit and Frommel, 1989; Gros *et al.*, 1989a; Gros, 1990): two high-affinity sites, Ca1 and Ca2, of about equal strength ($K_{diss} \approx 10^{-10}$ M), and a low-affinity site Ca3 (K_{diss} 10^{-4} to 10^{-7} M). Thermitase has all three Ca^{2+}-binding sites, the subtilisins have Ca1 and Ca3, and proteinase K has only Ca3. The occupancy of these sites depends on the calcium ion concentration in solution. The Ca1 site is the only one still fully occupied at 0 mM $CaCl_2$ (Gros *et al.*, 1991), while the Ca3 site is occupied by a Na^+ or K^+ ion at low Ca^{2+} concentration.

Based on sequence homology (Figure 3) and the known ligands of Ca^{2+} at these sites, we can now predict Ca^{2+}-binding sites in all of the other subtilases (Table VII). Ca1 and Ca2 sites, both in external loops, are relatively easy to predict since many side-chain carbonyl and/or carboxyl oxygens are required. The Ca1 site should be present in nearly all of the class I subtilases, but not in SEEPIP or EFCYLA and not in any of the class II subtilases, since they all lack the essential loop 76–81. The Ca2 site should definitely be present in at least five other subtilases than thermitase (viz. BSISP1, LLSK11, HSFURI, KLKEX1 and VAPROA), but in none of the class I-S subtilases (subtilisin branch). Various other candidates for Ca1 and Ca2 sites lack one or two side-chain ligands, but these may be provided by adjacent residues. In these cases Ca^{2+} binding may be strengthened through protein engineering by introducing Asp or Asn residues at the appropriate positions. For instance, an N76D substitution in subtilisin BPN′ improved the stability, presumably by strengthening Ca1 binding (Pantoliano *et al.*, 1989). It should even be possible to introduce the high-affinity Ca2 site into the class I-S and most of the class II subtilases.

The weak Ca3 site is difficult to predict since it has only two side-chain ligands, both from D197 in thermitase. The mutation D197E weakens or abolishes Ca^{2+} binding (Pantoliano *et al.*, 1988) and this substitution is found naturally in five subtilases, including subtilisin Carlsberg. However, this Ca3 site can be strengthened by the introduction of negatively charged side chains in the vicinity of the bound Ca^{2+}, e.g. G131D and/or P172D; the strongest Ca^{2+} binding at the Ca3 site was achieved by introducing the combination of D131, D172 and D197 (Pantoliano *et al.*, 1988). Therefore we have used as a criterion for the prediction of binding at Ca3 either the presence of D197, or of a combination of E197 with D131 and/or D172.

Additional Ca^{2+}-binding sites, not present in thermitase or in the subtilisins, may occur in the subtilases with insertions in VRs. Indeed many of these VRs are rich in Asp, Asn and Glu residues which could provide side-chain ligands.

Charged residues and ionic interactions

Charge is conserved at very few positions in all subtilases, and in fact only the positive charge on residue 94(K/R) and the negative charges on residues D32 (active site) and 41(D/E) are invariable. In class II subtilases charge is also totally conserved at R42, D60 and D197.

A number of ionic interactions has been located in subtilisin BPN′ and thermitase (Teplyakov *et al.*, 1990), and in proteinase K (Betzel *et al.*, 1988a,b,c, 1990). Very few of these are conserved amongst the four known structures (Table VIII). A search for such interacting charged residues at topologically equivalent positions in other subtilases indicates that their occurrence may be rather low, as shown in Table VIII, with the exception of the salt bridges that stabilize the Ca2 site (49–94, 52–94) and the Ca3 site (197–247). This finding is in agreement with observations in other protein families that conservation of ion-pairs is low, unless they have more specific functions to perform (Barlow and Thornton, 1983).

Introduction of ionic interactions, such as those in Table VIII, through protein engineering, may improve enzyme stability towards denaturants, alkaline pH or high temperature. The stabilizing mutation K213R (Cunningham and Wells, 1987) occurs only three times in natural sequences, while the destabilizing mutation K170E is not found at all (Figure 3).

Aromatic residues and interactions

Aromatic residues are moderately well conserved in all subtilases only at positions 6, 50, 113 and 189. Far more aromatic side chains are totally conserved in the class II subtilases: W6, Y15, Y29, F42, Y82, 113, 189 and 192.

Clustered aromatic residues, with potential interactions between them (Table VIII) have been located in thermitase (Teplyakov *et al.*, 1990). From the sequence alignment we predict that few

Table VII. Ca^{2+}-binding sites in subtilases

Class	Acronym	Ca1	Ca2	Ca3
I	BASBPN	●	–	●
	BSS168	○	–	○
	BSSDY	○	–	○
	BLSCAR	●	–	●
	BAPB92	○	–	○
	BYSYAB	○	–	○
	BLS147	○	–	–
	BSEPR	○	–	○
	BSISP1	?	○	○
	TVTHER	●	●	●
	BSBPF	?	–	–
	EFCYLA	–	?	–
	SEEPIP	–	–	○
	SPSCPA	?	–	○
	LLSK11	○	○	○
	DNEBPR	○	?	–
	XCEXPR	○	–	○
	SMEXSP	?	–	–
	AVPRCA	?	–	–
	MMPPC3	?	?	–
	HSIPC2	○	?	–
	HSFURI	○	○	–
	DMFUR1	○	?	–
	KLKEX1	○	○	–
	SCKEX2	○	?	–
II	VAPROA	–	○	○
	TRT41A	–	–	○
	TAAQUA	–	–	○
	TAPROK	–	–	●
	TAPROR	–	–	○
	TAPROT	–	–	○
	ACALPR	–	–	○
	AOALPR	–	–	○
	SCPRB1	–	–	○
	YLXPR2	–	?	○

●, known Ca^{2+}-binding sites from X-ray structures; ○ probable binding sites from sequence homology; ? uncertain binding sites; – no binding site.

Known main chain (m) and side chain (s) ligands for the Ca^{2+} ions from the X-ray structures are:

Ca1: 2Q/D(s), 41D(s,2x), 75(m), 77N/D(s), 79(m), 81(m).

Ca2: 49D/N(s), 52D/N(s), 54D(s,2x), 56(m), 58Q(s), H_2O(1x), and stabilized by 94R(s).

Ca3: 194(m), 197D(s,2x), H_2O(4x).

of the listed aromatic interactions are conserved in other subtilases, with the exception of the cluster at residues 48, 50, 91 and 113. Nevertheless, the introduction of aromatic residues at appropriate positions may enhance stability, as has been demonstrated for the M50F mutant of subtilisin BPN′ (Cunningham and Wells, 1987; Pantoliano *et al.*, 1989).

Random mutants

Through random mutagenesis studies several point mutants of subtilisin BPN′ have been identified with either enhanced thermal stability, i.e. S53T, A116E, L126I, S188P, Q206C, N218S, T254A and others (Bryan *et al.*, 1986; Rollence *et al.*, 1988), or enhanced alkaline stability, i.e. M50F, I107V and K213R (Cunningham and Wells, 1987). Combinations of these stabilizing mutations lead to a nearly additive increase in ΔG of unfolding. The substitution N218S increases thermostability by 4°C, due to slightly improved hydrogen bonding.

In the natural subtilases, S218 occurs most frequently, and N218 and T218 are the main alternatives (Figure 3). Interestingly, the thermostable enzyme aqualysin (TAAQUA) has retained the less favourable N218. At position 254 Ala is already the most frequently occurring residue, while at 107 both Ile and Val occur often. On the other hand, P188 is found only three times, the variants T53, E116 and I126 only once, and C206 does not occur at all.

Discussion

In the present study we have first defined the structurally conserved regions (SCRs) or three-dimensional framework of the catalytic domain of all subtilases, the family of subtilisin-like serine proteases. This framework contains 191 residues and, as shown schematically in Figure 6, consists in essence of an internal core of seven parallel β-sheet strands (e1–e7) and two buried helices (hC and hF), surrounded by five amphipathic helices (hA, hD, hE, hG and hH) and two anti-parallel β-sheet strands (e8 and e9). From the amino acid sequence alignment of 35 members of the subtilase family (Figure 3) we have identified the SCRs in each of these proteases. This knowledge provides the basis for building a three-dimensional model of any of the subtilases. Crude models of lactococcal proteinase LLSK11 (Vos *et al.*, 1991) and the class I-E subtilases, notably furin (van de Ven *et al.*, 1990) have been proposed using these data.

Subsequent steps in homology modelling are to predict the conformations of connecting backbone segments (VRs) between these secondary structure elements, and then the conformations

Table VIII. Ionic and aromatic interactions in subtilases

Interacting residues	Class I				Class II		Type of interaction
	BASBPN	BLSCAR	TVTHER	Other	TAPROK	Other	
Ionic:							
10–184	–	–	K-D	1	R-D	7	hA–t10
87– 22	–	–	K-E	3	–	0	
94– 49	–	–	R-D	10	–	3	Ca2 stability
94– 52	–	–	R-D	9	–	2	Ca2 stability
136–140	K-D	K-D	–	5	–	1	hE-hE
141–112	K-E	–	–	5	R-D	1	hD-hE
145–116	–	–	K-D	0	–	0	hD-hE
170–195	K-E	K-E	–	3	–	0	t8-t12
185–181	–	–	–	3	R-D	2	
247–197	R-D	R-E	R-D	11	–	1	Ca3 stability
247–251	R-E	–	R-E	1	–	0	hG-hG
267–184	–	–	R-D	1	–	0	
267–255	–	–	R-D	1	–	0	
272–255	–	–	K-D	1	–	0	
Aromatic:							
4–206	–	–	Y-Y	0	–	0	Nterm-e8
4–214	–	–	Y-Y	2	–	0	Nterm-e9
4[+1]– 17	–	–	F-W	10	–	0	Nterm-hB
48– 50	–	–	W-F	6	–	1	
48–113	–	–	W-Y	4	–	5	e2-hD
50–113	–	F-W	F-Y	7	Y-F	4	e2-hD
91–113	Y-W	Y-W	–	6	F-F	3	e3-hD
167–170	–	–	Y-Y	1	–	0	
167–171	Y-Y	Y-Y	Y-Y	8	–	0	
171–195	–	–	Y-W	0	–	0	t9-t12
192–262	–	–	Y-Y	1	–	1	
261–262	F-Y	F-Y	–	2	–	0	
262–263	Y-Y	Y-Y	Y-W	3	–	0	

Underlined interactions are known, and from these all other interactions are predicted at topologically equivalent positions. 'Other' indicates the number of other subtilases in which the interaction is predicted.

of side chains. Depending on whether the unknown subtilase belongs to class I or class II, the relevant information for these predictions can be obtained from the most similar known structure. For instance, since class II subtilases have relatively high sequence homology and low variation in the VR lengths, their model structures can be derived directly from the proteinase K structure. On the other hand, many of the class I subtilases have numerous inserts in the VRs, which make it impossible to model build them entirely from structural homology.

The most remarkable catalytic domains in terms of size are those of SPSCPA and LLSK11, which have 216 and 238 additional residues respectively, inserted in various VRs. The single inserts of 134 and 151 residues respectively, in VR13 are even large enough to be considered as separate (sub)domains. Protein folding is evidently not greatly affected by such large inserts, since the framework secondary structure elements are still able to find each other, despite the fact that they are now much farther apart in the primary sequence. It is also interesting to note that in the lactococcal proteinases (LLSK11 and variants) a much higher frequency of natural point mutations is found in the large insert VR13 than in the SCRs of the catalytic domain (Vos *et al.*, 1991). This implies that the large insert is under less evolutionary pressure for conservation of structure and function. However, the function of such large inserts and their evolutionary significance remains unclear.

Large inserts such as these found in subtilases have also been observed in other protein families, such as the aminoacyl−tRNA synthetases (Burbaum *et al.*, 1990). Protein-engineering studies have demonstrated that, at least in some proteins, large insertions or deletions can be made in surface loops without severely affecting protein function or stability (Kuipers *et al.*, 1989; Burbaum *et al.*, 1990; P.Vos *et al.*, in preparation). This phenomenon can be exploited by grafting loops or even domains with known properties, such as substrate or metal-ion-binding sites, into VRs of subtilases. A recent example of this type of protein engineering is the grafting of a Ca-binding loop of thermolysin into *Bacillus subtilis* neutral protease (Toma *et al.*, 1991).

Next, to be able to predict which residues in each of the subtilases contribute to catalytic activity, substrate binding and structural stability, we have combined information on sequence homology with the known interactions in the three-dimensional structures of our four reference proteins. This knowledge provides the basis for developing a protein-engineering strategy to modify one or more of the properties of these subtilases. After a brief review of the numerous engineering studies performed to date on subtilisin, we have provided similar and new suggestions for protein engineering of the other subtilases.

Important targets for engineering are the industrial proteases, and in particular the extracellular subtilases of *Bacillus* and *Tritirachium* used in detergent formulations, and those from *Lactococcus* and *Aspergillus* used in food processing. Main goals of protein engineering should be first to tailor the stability of these subtilases to withstand the often harsh industrial conditions of either temperature, pH, pressure, salinity, oxidizing agents or detergents, and secondly to obtain optimal substrate specificity and catalytic activity of these enzymes in specific environments.

In terms of stability, several other factors than the numerous hydrogen bonds and hydrophobic interactions may contribute to overall stability, such as Ca^{2+} ion binding, disulphide bonds, ionic and aromatic interactions. Clearly, different choices have been made during natural evolution of these subtilases to obtain the proper balance between activity and stability under various environmental conditions. In the thermostable enzymes proteinase K and aqualysin, the two S-S bonds in each structure appear to be responsible for extra stability. On the other hand, in the thermostable enzyme thermitase no S-S bonds are present, and extra stability is apparently achieved by extensive ionic and aromatic interactions, as well as binding of three Ca^{2+} ions (Tables VI−VIII). By these criteria, the proteases of *Staphylococcus* (SEEPIP), *Aspergillus* (AOALPR) and *Acremonium* (ACALPR) should be the least stable, since they are predicted to have only the weak Ca3 site and no S-S bonds. In contrast, the class I-E subtilases should be very stable, as we predict one or two S-S bonds in addition to the strong Ca1 and Ca2 sites.

It would appear that maximally two disulphide bonds are naturally present in the other known subtilases. However, this number could be increased by engineering of cysteines at appropriate sites as suggested (Table VI). All three known Ca^{2+}-binding sites in TVTHER are predicted to occur only in LLSK11 and possibly in BSISP1 (Table VII). The other subtilases may be further stabilized by the introduction of the missing Ca-binding sites, either by the suggested simple point mutations or by grafting of an entire Ca-binding loop. In general, stabilizing effects of single point mutations in subtilisin have been found to be additive (Rollence *et al.*, 1988; Pantoliano *et al.*, 1989). However, one should remember that several natural compensatory substitutions are known in subtilisins (McPhalen and James, 1988), for instance to occupy equivalent internal volumes. This compensatory effect must be taken into consideration in engineering strategy. Furthermore, one must consider that the selection or engineering for high stability can be offset by a lowered activity of the mutant protease.

In terms of substrate specificity and catalytic activity, we have identified the conserved as well as variable residues involved in substrate binding (Table V). Any variation in specificity of natural subtilases is most likely due to modulation of these variable residues. Engineering studies have demonstrated that the specificity of one subtilase can be completely transferred to a related subtilase by a very limited number of amino acid replacements (Wells *et al.*, 1987a; Vos *et al.*, 1991). We have provided a hypothesis for the high specificity for Arg/Lys P1 residues of some subtilases, and for the specificity for paired basic residues in the prohormone convertases (van de Ven *et al.*, 1990). We propose that substrate specificity may not only be modulated by natural variations in the amino acid sequence of segment 96−104, but also by variations in the length of this exposed loop (VR7). Indeed, deletion of residues 97−101 in BLSCAR abolishes activity (Schülein *et al.*, 1991) and our own engineering studies of LLSK11 have demonstrated that deletions, insertions and point mutations in this loop can alter specificity and catalytic activity (P.Vos, I.van Alan-Boerrigter, P.Bruinenberg, F.A. Exterkate, M.Nijhuis, W.M.de Vos and R.J.Siezen, in preparation).

Evidently there are numerous ways to modulate substrate binding and specificity in subtilases by protein engineering. It should be obvious though that since so many residues contribute to substrate binding, it may be difficult in most cases to predict the net effect of even a single point mutation in the binding region.

The family tree of the subtilase catalytic domains (Figure 4), as deduced from the amino acid sequences, does not represent the simple divergent evolution of a single gene. The ancestral subtilase gene must have been present very early in evolution, prior to the divergence of bacteria, archaea and eukaryotes (Woese *et al.*, 1990). Duplication of this ancestral subtilase gene,

followed by an independent evolution, could explain the presence of subtilases of both classes in Gram-negative bacteria and yeasts. Further duplications must have taken place in later evolution, since at least three distinctly different and specialized subtilases are found in present-day species such as *B.subtilis* and man.

In some branches of the family tree the catalytic domain gene was coupled to genes coding for different functions, most notably membrane anchoring (Table I). In the class I-E subtilases this fusion of domains took place prior to the divergence of lower and higher eukaryotes, because considerable sequence homology also exists in parts of their C-terminal extensions (van de Ven *et al.*, 1990). Such an extended homology in C-terminal domains also applies for the lactococcal (LLSK11) and streptococcal (SPSCPA) subtilases (R.J.Siezen and W.M.de Vos, in preparation), but not for the related subtilases from *Vibrio* (VAPROA) and *Thermus* (TAAQUA), nor for the multiple subtilases from *Bacillus* (BSBPF, BSEPR). In the superfamily of trypsin-like serine proteases additional domains have also been added in several cases, both at the N- and C-terminal sides of the catalytic domain, through an evolutionary process called exon shuffling (Doolittle, 1985, 1989; Rogers, 1985; Irwin *et al.*, 1988). This evolutionary gene or domain coupling process may also apply to the subtilases, but the knowledge of gene intron−exon structure in this family is presently limited to HSFURI (van de Ven *et al.*, 1990), TAPROK (Gunkel and Gassen, 1989), TAPROT (Samal *et al.*, 1989) and AOALPR (Cheevadhanarah *et al.*, 1991), and therefore insufficient for analysis of exon shuffling.

Finally, the recent discovery by Tomkinson and Jonsson (1991) of HSTPP, a subtilase family member with a modified function (i.e. exopeptidase rather than endopeptidase) increases the likelihood that subtilases with entirely different, possibly even non-enzymatic functions may have evolved, as previously found in other protein families (Maeda *et al.*, 1984; Huber and Carrell, 1989; Doolittle, 1989).

Acknowledgements

We are greatly indebted to our colleagues Drs W.J.M.van de Ven, A.J.M.Roebroek, M.Booth, A.P.Sloma, B.Samal, G.Engelke, A.Kortt, D.Godette and J.C.van der Laan for communicating their sequence data prior to publication. We thank our colleagues Drs S.Visser, F.A.Exterkate, P.Bruinenberg and O.Kuipers for critically reading this manuscript. Use of the services and facilities of the Dutch National NWO/SURF Expertise Center CAOS/CAMM, under grant numbers SON326-052 and STW NCH99.1751, is gratefully acknowledged. This work was supported in part by contract BAP-0011-NL of the Commission of European Communities.

References

Abad-Zapatero,C., Rydel,T.J. and Erikson,J. (1990) *Proteins Struct. Funct. Genet.*, **8**, 62−81.
Barlow,D.J. and Thornton,J.M. (1983) *J. Mol. Biol.*, **168**, 867−885.
Bazan,J.F. and Fletterick,R.J. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 7872−7876.
Betzel,C., Belleman,M., Pal,G.P., Bajorath,J., Saenger,W. and Wilson,K.S. (1988a) *Proteins Struct. Funct. Genet.*, **4**, 157−164.
Betzel,C., Pal,G.P. and Saenger,W. (1988b) *Acta Crystallogr., B*, **44**, 163−172.
Betzel,C., Pal,G.P. and Saenger,W. (1988c) *Eur. J. Biochem.*, **178**, 155−171.
Betzel,C., Teplyakov,A.V., Harytyunyan,E.H., Saenger,W. and Wilson,K.S. (1990) *Protein Engng*, **3**, 161−172.
Blundell,T.L., Jenkins,J.A., Sewell,B.T., Pearl,L.H., Cooper,J.B., Tickle,I.J., Veerapandian,B. and Wood,S.P. (1990) *J. Mol. Biol.*, **211**, 919−941.
Bode,W., Papamokos,E., Musil,D., Seemuller,U. and Fritz,H. (1986) *EMBO J.*, **5**, 813−818.
Bode,W., Papamokos,E. and Musil,D. (1987) *Eur. J. Biochem.*, **155**, 673−692.
Bott,R., Ultsch,M. Kossiakoff,A., Graycar,T., Katz,B. and Power,S. (1988) *J. Biol. Chem.*, **263**, 7895−7906.
Briedigkeit,L. and Frommel,C. (1989) *FEBS Lett.*, **253**, 83−87.
Bryan,P.N., Rollence,M.L., Pantoliano,M.W., Wood,J., Finzel,B.C., Gilliland,G.L., Howard,A.J. and Poulos,T.L. (1986) *Proteins Struct. Funct. Genet.*, **1**, 326−334.
Burbaum,J.J., Starzyk,R.M. and Schimmel,P. (1990) *Proteins Struct. Funct. Genet.*, **7**, 99−111.
Carter,P. and Wells,J.A. (1990) *Proteins Struct. Funct. Genet.*, **7**, 335−342.
Cheevadhanarah,S., Saunders,G., Renno,D.V., Holt,G. and Flegel,T. (1991) Accession code X54726, EMBL Data Library.
Chen,C.C. and Cleary,P.P. (1990) *J. Biol. Chem.*, **265**, 3161−3167.
Chothia,C. and Lesk,A.M. (1986) *EMBO J.*, **5**, 823−829.
Cunningham,B.C. and Wells,J.A. (1987) *Protein Engng*, **1**, 319−325.
Doolittle,R.F. (1985) *Trends Biochem. Sci.*, **10**, 233−237.
Doolittle,R.F. (1989) *Trends Biochem. Sci.*, **14**, 244−245.
Eijsink,V.G.H., Vriend,G., van de Burg,B., Venema,G. and Stulp,B.K. (1990) *Protein Engng*, **4**, 99−104.
Estell,D.A., Graycar,T.P. and Wells,J.A. (1985) *J. Biol. Chem.*, **260**, 6518−6521.
Estell,D.A., Graycar,T.P., Miller,J.V., Powers,D.B., Burnier,J.P., Ng,P.G. and Wells,J.A. (1986) *Science*, **233**, 659−663.
Exterkate,F.A., Alting,A.C. and Slangen,C.J. (1991) *Biochem. J.*, **273**, 135−139.
Felsenstein,J. (1990) *PHYLIP Manual Version 3.3*. University Herbarium, University of California, Berkeley, CA, USA.
Fitch,W.M. and Margoliash,E. (1967) *Science*, **155**, 279−284.
Frommel,C. and Sander,C. (1989) *Protein Struct. Funct. Genet.*, **5**, 22−37.
Fuller,R.S., Brake,A. and Thorner,J. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 1434−1438.
Gorbalenya,A.E., Donchenko,A.P., Blinov,V.M. and Koonin,E.V. (1989) *FEBS Lett.*, **243**, 103−114.
Greer,J. (1981) *J. Mol. Biol.*, **153**, 1027−1042.
Greer,J. (1990) *Protein Struct. Funct. Genet.*, **7**, 317−334.
Gros,P., Kalk,K.H. and Hol,W.G.J. (1991) *J. Biol. Chem.*, **266**, 2953−2961.
Gros,P., Betzel,Ch., Dauter,Z., Wilson,K.S. and Hol,W.G.J. (1989a) *J. Mol. Biol.*, **210**, 347−367.
Gros,P., Fujinaga,M., Dijkstra,B.W., Kalk,K.H. and Hol,W.G.J. (1989b) *Acta Crystallogr., B*, **5**, 488−499.
Gros,P., van Gunsteren,W.F. and Hol,W.G.J. (1990) *Science*, **249**, 1149−1152.
Gunkel,F.A. and Gassen,H.G. (1989) *Eur. J. Biochem.*, **179**, 185−194.
Hazes,B. and Dijkstra,B.W. (1988) *Protein Engng*, **2**, 119−125.
Higgins,D.G. and Sharp,P.M. (1988) *Gene*, **73**, 237−244.
Hirono,S., Akagawa,H., Mitsui,Y. and Iitaka,Y. (1984) *J. Mol. Biol.*, **178**, 389−413.
Huber,R. and Carrell,R.W. (1989) *Biochemistry*, **28**, 8951−8966.
Irwin,D.M., Robertson,K.A. and MacGillivray,R.T.A. (1988) *J. Mol. Biol.*, **200**, 31−45.
Kabsch,W. (1976) *Acta Crystallogr., A*, **32**, 922−923.
Kabsch,W. and Sander,C. (1983) *Biopolymers*, **22**, 2577−2637.
Katz,B.A. and Kossiakoff,A. (1987) *J. Biol. Chem.*, **261**, 15480−15485.
Katz,B. and Kossiakoff,A.A. (1990) *Proteins Struct. Funct. Genet.*, **7**, 343−357.
Kim,D., Yang,Ch. and Choi,M. (1990) *Korean Biochem. J.*, **23**, 58−61.
Kuipers,O.P., Thunnissen,M.M.G.M., De Geus,P., Dijkstra,B.W., Drenth,J., Verheij,H.M. and De Haas,G.H. (1989) *Science*, **244**, 82−85.
Kwon,S.-T., Terada,I., Matsuzawa,J. and Ohta,T. (1988) *Eur. J. Biochem.*, **173**, 491−497.
Maeda,N., Yang,F., Barnett,D.R., Bowman,B.H. and Smithies,O. (1984) *Nature*, **309**, 131−135.
Masazumi,M., Signor,G. and Matthews,B.W. (1989) *Nature*, **342**, 291−293.
McPhalen,C.A. and James,M.N.G. (1988) *Biochemistry*, **27**, 6582−6598.
McPhalen,C.A., Svendsen,I., Jonassen,I. and James,M.N.G. (1985) *Proc. Natl Acad. Sci. USA*, **82**, 7242−7246.
Mitchinson,C. and Wells,J.A. (1989) *Biochemistry*, **28**, 4807−4815.
Mizuno,K., Nakamura,T., Ohshima,T., Tanaka,S. and Matsuo,H. (1988) *Biochem. Biophys. Res. Commun.*, **156**, 246−254.
Ogrydziak,D.M. (1988) *CRC Crit. Rev. Biochem.*, **8**, 177−187.
Pantoliano,M.W., Ladner,R.C., Bryan,P.N., Rollence,M.L., Wood,J.F. and Poulos,T.L. (1987) *Biochemistry*, **26**, 2077−2082.
Pantoliano,M.W., Whitlow,M., Wood,J.F., Rollence,M.L., Finzel,B.C., Gilliland,G.L., Poulos,T.L. and Bryan,P.N. (1988) *Biochemistry*, **27**, 8311−8317.
Pantoliano,M.W., Whitlow,M., Wood,J.F., Dodd,S.W., Hardman,K.D., Rollence,M.L. and Bryan,P.N. (1989) *Biochemistry*, **28**, 7205−7213.
Pearson,W.R. and Lipman,D.J. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 2444−2448.
Prager,E.M. and Wilson,A.C. (1978) *J. Mol. Evol.*, **11**, 129−142.
Riisgaard,S. (1990) In Christiansen,C., Munck,L. and Villadsen,J. (eds), *Proc. 5th Eur. Congr. Biotech.*, Munksgaard, Copenhagen, Vol. 1, pp. 31−40.
Robertus,J.D., Kraut,J., Alden,R.A. and Birktoft,J.J. (1972) *Biochemistry*, **11**, 4293−4303.
Rogers,J. (1985) *Nature*, **315**, 458−459.

Rollence,M.L., Filpula,D., Pantoliano,M.W. and Bryan,P.N. (1988) *CRC Crit. Rev. Biotech.*, **8**, 217–224.
Russell,A.J. and Fersht,A.R. (1987) *Nature*, **328**, 496–500.
Russell,A.J., Thomas,P.G. and Fersht,A.R. (1987) *J. Mol. Biol.*, **193**, 803–813.
Saitou,N. and Nei,M. (1987) *Mol. Biol. Evol.*, **4**, 406–425.
Samal,B.B., Karan,B., Boone,T.C., Chen,K.K., Rohde,M.F. and Stabinsky,Y. (1989) *Gene*, **85**, 329–333.
Schechter,I. and Berger,A. (1967) *Biochem. Biophys. Res. Commun.*, **27**, 157–162.
Schülein,R., Kreft,J., Gonski,S. and Goebel,W. (1991) *Mol. Gen. Genet.*, **227**, 137–143.
Signor,G., Vita,C., Fontana,A., Frigerio,F., Bolognesi,M., Toma,S., Gianna,R., De Gregoris,E. and Grandi,G. (1990) *Eur. J. Biochem.*, **189**, 221–227.
Smeekens,S.P. and Steiner,D.F. (1990) *J. Biol. Chem.*, **265**, 2997–3000.
Sneath,P.H.A. and Sokal,R.R. (1973) *Numerical Taxonomy*. W.H.Freeman & Co., San Francisco, pp. 230–234.
Sobek,H., Hecht,H.J., Hofman,B., Aehle,W. and Schomburg,D. (1990) *FEBS Lett.*, **274**, 57–60.
Sutcliffe,M.J., Haneef,I., Carney,D. and Blundell,T.L. (1987a) *Protein Engng*, **1**, 377–384.
Sutcliffe,M.J., Hayes,F.R.F. and Blundell,T.L. (1987b) *Protein Engng*, **1**, 385–392.
Takagi,H., Morinaga,Y., Ikemura,H. and Inouye,M. (1988) *J. Biol. Chem.*, **263**, 19592–19596.
Takagi,H., Takahashi,T., Momose,H., Inouye,M., Maeda,Y., Matsuzawa,H. and Ohta,T. (1990) *J. Biol. Chem.*, **265**, 6874–6878.
Teplyakov,A.V., Kuranova,I.P., Harutyunyan,E.H. and Vainshtein,B.K. (1990) *J. Mol. Biol.*, **214**, 261–279.
Terada,I., Kwon,S.-T., Miyata,Y., Matsuzawa,H. and Ohta,T. (1990) *J. Biol. Chem.*, **265**, 6576–6581.
Toma,S., Campagnoli,S., Margarit,I., Gianna,R., Grandi,G., Bolognesi,M., Filippis,V.De and Fontana,A. (1991) *Biochemistry*, **30**, 97–106.
Tomkinson,B. and Jonsson,A.-K. (1991) *Biochemistry*, **30**, 168–174.
van de Ven,W.J.M., Voorberg,J., Fontijn,R., Pannekoek,H., van den Ouweland,A.M.W., van Duijnhoven,H.L.P., Roebroek,A.J.M. and Siezen,R.J. (1990) *Mol. Biol. Rep.*, **14**, 265–275.
Vos,P., Simons,G., Siezen,R.J. and de Vos,W.M. (1989) *J. Biol. Chem.*, **264**, 13579–13585.
Vos,P., Boerrigter,I.J., Buist,G., Haandrikman,A.J., Nijhuis,M., de Reuver,M.B., Siezen,R.J., Venema,G., de Vos,W.M. and Kok,J. (1991) *Protein Engng*, **4**, 479–484.
Weber,I.T. (1990) *Proteins Struct. Funct. Genet.*, **7**, 172–184.
Wells,J.A. and Estell,D.A. (1988) *Trends Biochem. Sci.*, **13**, 291–297.
Wells,J.A. and Powers,D.B. (1986) *J. Biol. Chem.*, **261**, 6564–6570.
Wells,J.A., Cunningham,B.C., Graycar,T.P. and Estell,D.A. (1987a) *Proc. Natl Acad. Sci. USA*, **84**, 5167–5171.
Wells,J.A., Powers,D.B., Bott,R.R., Graycar,T.P. and Estell,D.A. (1987b) *Proc. Natl Acad. Sci. USA*, **84**, 1219–1223.
Woese,C.R., Kandler,O. and Wheelis,L. (1990) *Proc. Natl Acad. Sci. USA*, **87**, 4576–4579.
Yanagida,N., Uozumi,T. and Beppu,T. (1986) *J. Bacteriol*, **166**, 937–944.

Received on May 7, 1991; accepted on July 30, 1991

APPENDIX A:

Naturally occurring amino acids, and their characteristics, at topologically equivalent positions in subtilases, compiled from sequences in Fig. 3 only.

res: residue number; **C:** subtilase class I or both classes (blank); **ss:** secondary structure (see Fig.2 and text); **core:** defined in Fig.2 and text; **acc:** average solvent accessibility in the four reference structures, (0) 0 Å^2, (-) 0–10 Å^2, (+-) 10–20 Å^2, (+) 20–40 Å^2, (++) >40 Å^2, (v) very variable; **invar:** totally invariant residues (100% conserved) and in brackets semi-invariant residues (>80% conserved); **main:** main residues observed at least four times, in order of decreasing frequency (underlined: >50%); **other:** type or number of different residues observed 1-3 times, in decreasing order; **characteristics:** see text for details.

				observed residues			
res	C	ss	core	acc	invar	main	other	characteristics
2	I			v		D,Q	T,R	Ca1 ligand (s)
5				+-		P,Q	7	
6		hA		++		W,Y	N,E,A	aromatic
7		hA		-		G,Y,H	6	
8		hA		-		L,I	5	nonpolar
9		hA		++		S,A,D	10	
10		hA		v		R,N,Q	10	
11		hA		o		L,V,T	4	nonpolar
12				++		S,K,Q,D,N	5	polar
13	I	hB		o		A,V	4	
14	I	hB		++		P	13	
15	I	hB		++		A,K	9	
16	I	hB		-		A,L,V	S,D,M	nonpolar
17	I	hB		+		W,H,Q	Y,P,N	
21	I			v		Y,I	7,-	nonpolar
22				v		T,K,A,D	4	
23				o	(G)	G	R	abnormal tors.angles
24		t1		++		K,S,A,T	6	polar
25		t1		++		G,N	D,H,K	
26				-		V,T,I	A,S	
27		e1		++		V,K,T	6	disulfide(to 117+[11])
28		e1		o		V,I,A	S	nonpolar
29		e1		o		A,Y	G,S,T	nonpolar
30		e1		-		V,I	L,S	nonpolar
31		e1		o		I,V,L	M	nonpolar
32		e1		-	D			active site
33				-		T,S,D	A,N	polar
34				-	(G)	G	A	abnormal tors.angles
35				-		I,V,L	C,F	nonpolar
36				+		D,E,N,L	6	polar
37		t2		+		-,S,T,A	10	
38		t2		++		S,N,T,D	4	polar
39				-	(H)	H	N	positive
40		t3		++		P,E,T	8	
41		t3		o		D,E	A,N	Ca1 ligand (s);negative
42				-		L,F	I,M,W	nonpolar
43				v		R,N,K,D	5	polar
44		e2		+-		A,V,I,P	4	nonpolar
45		e2		++		V,E,A,R,G	8	
46		e2		+		G,A	11	
47		e2		+		G,S	6	
48		e2		++		Y,A,K,F	W,H,R	
49		e2		+		D,S,N	T,A,R	Ca2 ligand (s)
50				v		F,A,Y	S	aromatic
51	I			v		V,N,S	4	
52	I			++		D,S,P	6	Ca2 ligand (s)
59				++		Q,T,L,A	10	Ca1 ligand (s)
60				+-		D,N	5	polar
61		s1		++		G,D,S	9	disulfide (to 98)
62		s1		++		N	9	polar
63		s1		-		G	10	
64		hC		-	H			active site
65		hC		o	G			
66		hC		o	(T)	T	M,S	polar
67		hC		o		H,R	Q,L,A	positive
68		hC		o		V,C		nonpolar
69		hC		o	(A)	A	S,T	
70		hC		o	G			
71		hC		-		T,E,I	V,Q	
72		hC		o		I,V	A,M,L	nonpolar
73		hC		-		A,G	3,-	
74				o		A,G	S,M,-	
75				++		L,K,V,N	6,-	Ca1 ligand (m)
76	I			++		N,D,-	4	
77	I	s2		++		N	4,-	Ca1 ligand (s)
78	I	s2		++		S,G,N	4,-	
79	I	s2		++		I,T	6,-	Ca1 ligand (m)
80	I	s2		-		G,C	3,-	disulfide (to 213)?
81	I			-		V,G	6,-	Ca1 ligand (m)
82				v		Y,L,V,A	6	nonpolar
83				o	(G)	G	T	
84		s3		o	(V)	V,I	T,A	nonpolar
85		s3		o	(A)	A	5	nonpolar
86		t4		+-		P,K,Y	A,F	
87		t4		++		N,G,K,E,S	A,D,Q	polar
88				-		A,V,T	S	
89		e3		+		S,K,Q,N	4	polar
90		e3		-		L,I,V	W,M	nonpolar
91		e3		v		Y,L,V,I	7	nonpolar
92		e3		o		A,G,P	4	
93		e3		o		V,I	M,Y,T	nonpolar
94		e3		-		K,R		salt bridge to D49,D53
95				o		V,M	4	nonpolar
96				+-		L,F	3,-	P4,P2 subst.binding (s)
97				++		D,G,S	5,-	
98		t5		++		-,S,A	9	disulfide (to 61)
99		t5		++		-,N,S,G	3	
100				+		G	11,-	P4,P3,P2 subst.binding
101				++		S,D	7,-	P4,P3 subst.binding
102				+		G,T	9,-	P4,P3 subst.binding
103				++		T,S,Q	6	polar
104		hD		++		Y,D,T,L	7	P4 subst.binding
105		hD		++		S,A	6	polar
106		hD		v		D,G,S,W	6	P3 subst.binding
107		hD		+-		I,V,E	4	P4 subst.binding
108		hD		+		A,I,L,V	4	nonpolar
109		hD		++		A,S,Q,D,N	I,R,E	polar
110		hD		o		G,S,A	E	
111		hD		o		I,L,M,F	V,G	nonpolar
112		hD		++		E,D,I	8	
113		hD		++		W,Y,F,D	L,H,Q	aromatic
114		hD		o		A,V,S	N,G,M	nonpolar
115		hD		+		A,T,I,L,P	3,-	
116		hD		++		N,Q,A	7,-	polar
117		t6		++		N,D,-	6	
118		t6		v		G,K,N	9	
119	I			-		M,A	5	
120		e4		++		D,H,A	7	
121		e4		o		V,I	A,T	nonpolar
122		e4		o		I,A,Y,V	L	nonpolar
123		e4		-		N,S		polar
124				o		M,A	5	nonpolar
125				-	S			P2,P1 subst.binding
126				+		L,W	F,I,G	nonpolar; P4,P2 binding
127				++	G			P4,P3,P1 subst.binding
128				+		G,P,S	N,I,T	
129				++		G,P,T	6	
130				++		S,D,Y	8	
131				+		S,G,D	4	
132	I			v		S	9	
133		hE		++		T,S,D,A	7	polar
134		hE		++		A,L,S,T	7	
135		hE		-		L,V,A	7	nonpolar
136		hE		++		K,D,E,N,Q	5	polar
137		hE		++		Q,K,S	7	polar
138		hE		-	(A)	A	4	nonpolar
139		hE		o		V,F,I	4	nonpolar
140		hE		++		N,D,A	9	
141		hE		++		N,K,Y,R	7	
142		hE		o		A,G	6	
143		hE		++		V,T,Y	7	
144		hE		++		S,A,N,E,K	Q,T,R	polar
145		t7		++		S,K,A,Q,R	5	polar
146		t7		+	(G)	G	D,N	
147				-		V,I,S	5	
148		e5		-		L,V,I	6	nonpolar
149		e5		-		V,F,I,Y	4	nonpolar
150		e5		o		V,A	4	nonpolar
151		e5		o		A,V,F,W	5	nonpolar
152		e5		-	(A)	A,S		P1 subst.binding
153				-		A,S	V	
154				+-	G			P1 subst.binding;abnor.φ,
155				++	(N)	N	D	oxyanion, P1,P1' binding
156				++		D,E,G,S	4	P1 subst.binding
157				v		G,N	A,S,D	
166				v		G,Y,S	5	P1 subst.binding
167				++		Y,S	6	aromatic cluster
168		t8		-		P,T	A	
169		t8		-		A,N,G	S,D	P1 subst.binding
170		t8		++		S,R,K,N	5	polar
171				+		Y,I,A,E	5	aromatic cluster
172		t9		++		A,P,S,D	8	
173		t9		v		S,N,-	6	
174				-		V,A,T	4	
175		e6		-		I,L,M	4	nonpolar;disulfide(247)
176		e6		-		T,A,S	I	
177		e6		o		V,I	T	nonpolar
178		e6		o		G,A,S	N,T	
179		e6		o		A,S	N	
180		e6		o		T,V,I,A	4	
181				v		D,T,S	6	polar
182		t10		++		S,Q	11	polar
183		t10		++		N,S	8	polar
184				v		D,G,N	5	
185				++		N,Q,K,A	7	
186				++		R,L	8	
187				-		A,S,P		
188		t11		++		S	10	
189		t11		++		F,Y	W,S	aromatic
190				o		S	A,D,T	polar
191				-		N,E,S	5	polar
192				v		Y,F	8	
193				-		G,C	N	disulfide (to 163/164)?
194		t12		++		S,A,P	7	
195		t12		++		E,V,S,C	7	disulfide (to 163)
196				-		V,L,I,T	P,Y	nonpolar
197				+		D,E,L	3	Ca3 ligand (s)
198		e7		-		V,I,L,A	F,C	nonpolar
199		e7		o		F,S,M,V,A	T,L,Y	nonpolar
200		e7		o		A,T	G,F,S	
201		e7		o		P,Y	F,I	
202				o		G,S	A	
203		s4		+		S,V,T	6	
204		s4		++		S,N,D	4	polar
205		e8		-		I,V	Y	nonpolar
206		e8		++		L,Y,Q	10	
207		e8		o	(S)	S,T		polar
208		e8		o		T,A,S		polar
209		e8		v		W,Y,D,L,V	S,Q,I	
210		t13		+		P,L,I,Y	7	
211		t13		++		G,N,D,S	6	
212		t13		++		N,S,G	7	polar
213		e9		++		C,T,G,A,S	K,R,Y	disulfide (to 80)?
214		e9		++		Y,T	S	
215		e9		v		A,N,G	8	
216		e9		++		S,T	10	polar
217		e9		++		L,I,H,Y	6	P1' subst.binding
218				++		S,N,T	4	P1' subst.binding
219				-	G			P1 subst.binding;abnor.φ,
220		hF		-	T			P1 subst.binding
221		hF		+-	S			active site
222		hF		-		M,A	L,Q,S	nonpolar; P1' binding
223		hF		o	(A)	A,S		
224		hF		o		T,A,S	C,G	
225		hF		o	P			
226		hF		-		H,L	4	
227		hF		o		V,A	I	nonpolar
228		hF		o		A,S	T,V	
229		hF		o	(G)	G	A	
230		hF		o		V,A,L,I	M,S	nonpolar
231		hF		o		A,L	8	nonpolar
232		hF		o		A	4	
233		hF		-		L,Y	A,V	nonpolar
234		hF		-		L,I,V,Y	4	nonpolar
235		hF		+-		L,K,M	Q,I,A	
236		hF		+		S,Q,E	5	polar
237				++		A,K,L,R	7	
239				++		P	11	
240				++		S,N,T,-	6	
242				++		T,S	5	polar
243		hG		+-		P,W,A,N	6	
244		hG		++		S,A,D	5	
245		hG		++		Q,D,E	6	polar
246		hG		-		V,L,I,M,A	T	nonpolar
247		hG		+		R,Q,E,A	6	salt bridge;disulfide
248		hG		++		A,D,N,H	8	
249		hG		v		L,R,I	9	
250		hG		-		L,I	5	nonpolar
251		hG		+		V,K,I	9	
252		hG		++		N,S,E	9	polar
253		hG		+		T,S	6	polar
254				+-		A,S	6	
255				v		T	10	polar
256				++		K,P,Y	10	
257	I			++		L,I	5	nonpolar
258	I			v		G,N	10	
262	I			++		Y,L	8	
263	I			v		Y,F	5	aromatic
264	I			v		G	A,K	
265	I			++		K,Y	7	
266				v		G,L	4	
267				-		L	7	nonpolar
268				-		V,L,I,A	M	nonpolar
269				+		N,Y,D	H,Q,R	
270	I	hH		o		A,V	L,I,Y	nonpolar
271	I	hH		++		K,E	10	
272	I	hH		v		A,K	4,-	
273	I	hH		-		A	5,-	nonpolar
274	I	hH		-		V,A	4,-	
275	I			++		Q	9,-	